Information Asymmetries and the Role of Information Intermediaries on Capital Markets

Dissertation zur Erlangung des wirtschaftswissenschaftlichen Doktorgrades der Wirtschaftswissenschaftlichen Fakultät der Georg-August-Universität Göttingen

Vorgelegt von: Steffen Bankamp, M.Sc.

Göttingen, Juni 2022

Betreuungsausschuss

Erstbetreuer	Prof. Dr. Jan Muntermann
Zweitbetreuer	Prof. Dr. Olaf Korn
Drittbetreuer	Prof. Dr. Lutz M. Kolbe

Danksagung

Diese Dissertation entstand während meiner Zeit an der Professur für Electronic Finance und Digitale Märkte der Georg-August-Universität Göttingen. Diese Zeit hat mich sehr geprägt und ich konnte mich fachlich und persönlich weiterentwickeln. Ein besonderer Dank gilt meinem Betreuer Prof. Dr. Jan Muntermann, der durch seine Unterstützung maßgeblich zu meiner Forschung beigetragen hat. Ich erinnere mich dabei an viele fruchtbare Diskussionen mit Prof. Muntermann, die mir sehr geholfen haben meine Forschungsarbeiten zu konzeptionieren und weiterzuentwickeln. Dabei hat er mir stets Orientierung vermittelt und mich gleichzeitig dazu motiviert, jenen Themengebieten und Fragestellungen zu folgen, die mich selber besonders interessieren.

Mein Dank gilt auch meinen Betreuern Prof. Dr. Olaf Korn und Prof. Dr. Lutz Kolbe, die mir in Gesprächen und bei Doktorandenkolloquien sehr gute Anregungen zu meiner Forschung gegeben haben.

Die Arbeit am Lehrstuhl ist immer auch eine Teamarbeit. Deshalb gilt meinen ehemaligen Kolleginnen und Kollegen Dennis Metzler, Nicole Neuss, Matthias Palmer, Jan Röder, Albert Torno, Stephanie Wurdak und den studentischen Hilfskräften ein ganz besonderer Dank. In der Lehre sowie in der Forschung konnte ich immer auf die Unterstützung des Lehrstuhlteams zählen. Ich erinnere mich immer gerne an die gemeinsamen Forschungsprojekte und an die zahlreichen fachlichen und persönlichen Gespräche. Besonders freut mich, dass aus dieser vertrauten Zusammenarbeit Freundschaften entstanden sind.

Ein ganz besonderer Dank gilt meinen Eltern, Großeltern und meiner Freundin Carla. Diese haben durch ihre Unterstützung, Motivation und Rücksichtnahme in den letzten Jahren ganz entscheidend zum Abschluss dieser Arbeit und zu meinem gesamten Studium beigetragen. Herzlichen Dank!

Table of Content

List of Figures .... **vi**

List of Tables .... **viii**

Abbreviations .... **x**

Symbols .... **xii**

A. Foundations **1**

1 Introduction 2
1.1 Motivation 2
1.2 Research Questions 3

2 Structure of the Thesis 5

3 Research Background 7
3.1 Theoretical Background 7
3.1.1 Information Asymmetries 7
3.1.2 Information Intermediaries 8
3.1.3 Signaling Theory 10
3.2 Related Literature 11
3.2.1 Natural Language Processing of Financial Documents 11
3.2.2 Information Intermediaries and Regulatory Change 12
3.2.3 Management’s Signaling Toward Capital Markets 14
3.3 Methods 15
3.3.1 Natural Language Processing 15
3.3.2 Difference-in-Differences Analysis 19
3.4 Datasets 21

B.Studies: Individual Research Contributions **24**

I. Research Area: Methodological Foundation **25**

I.1. Document Similarity **26**

1 Introduction 27

2 Theoretical Background 28
2.1 Document Similarity 28
2.2 Document Representations 30
2.2.1 Bag-of-Words Models 30
2.2.2 Topic Models 30
2.2.3 Word and Document Embeddings 31

2.2.4 Transformer-Based Models 31
2.3 Document Representation for Similarity Measurement in the Finance and Accounting Literature 32

3 Data 33

4 Experimental Design 34

5 Experimental Results 37
5.1 Three Dimensions of Similarity 37
5.2 Detailed Analysis of Object Similarity 39
5.3 Detailed Analysis of Temporal Similarity 41

6 Discussion 42

7 Conclusion 44

II. Research Area: Information Asymmetries and Information Intermediaries on Capital Markets 46

II.1. Information Extraction on Sustainability 48

1 Introduction 49

2 Research Background on Sustainability in Investing 50

3 Methodology and Problem Description 51
3.1 Design Science Research 51
3.2 Research Process 51
3.3 Problem Description 52

4 Artifact Design 53
4.1 Design Requirements 53
4.2 Design Principles 54
4.3 Design Features 55

5 Artifact Evaluation 57
5.1 Dataset 57
5.2 Evaluation Results 58

6 Discussion 64

7 Conclusion 65

II.2. MiFID II and Analyst Reports 66

1 Introduction 67

2 MiFID II, Research Services, and Cost Transparency 69

3 The Impact of MiFID II on Analyst Behavior 70

4 Methodology 72
4.1 Coverage 73
4.2 Optimism 74
4.3 Novelty 75

5 Dataset 77

6 Results 78
6.1 Coverage 78
6.2 Optimism 80
6.3 Novelty 82
6.4 Robustness Checks 86

7 Discussion 87

8 Conclusion 91

Appendix A – Variable Definition 92

Appendix B – Preprocessing of Analyst Reports 93

II.3. MiFID II and Media Coverage 94

1 Introduction 95

2 Institutional Background and Hypothesis Development 97
2.1 Institutional Background on MiFID II 97
2.2 MiFID II and its Effect on Analyst Research Output 98
2.3 Relation Between Analyst Coverage and News Coverage 99

3 Research Design 101
3.1 MiFID II and Analyst Coverage 103
3.2 Analyst Coverage Loss and Media Coverage 104
3.3 MiFID II and Media Coverage 105

4 Sample Selection and Descriptive Statistics 105

5 Results 107
5.1 MiFID II and Analyst Coverage 107
5.2 Analyst Coverage Loss and Media Coverage 109
5.3 MiFID II and Media Coverage 110
5.4 Effect Channel and Robustness Checks 113

6 Discussion 115

7 Conclusion 117

Appendix – Variable Definition 118

II.4. Signaling Digital Transformation 119

1 Introduction 120

2 Theoretical Foundations 123

2.1 The CDO as the Centralized Digital Transformation Responsibility in Firms 123

2.2 Signaling Theory and the Reduction of Information Asymmetries Through External Corporate Communication 125

2.3 Hypothesis Development 127

3 Methodological Approach 129

4 Datasets and Descriptive Statistics 131

5 Empirical Results 136

6 Discussion 138

6.1 Theoretical and Practical Implications 138

6.2 Limitations and Future Research 140

7 Conclusion 141

C. Contributions 143

1 Summary of Results 144

2 Implications 146

3 Limitations 148

4 Future Research 150

References 152

Appendix xiii

List of Figures

Figure 1. Strategies for Reducing Information Asymmetries 2

Figure 2. Overview of the Thesis 5

Figure 3. Overview of Information Asymmetries, Their Consequences, and Their Mitigations 7

Figure 4. Text Pre-Processing 15

Figure 5. Illustration of a Term-Document Matrix 17

Figure 6. Illustration of Document Similarity Calculations 19

Figure 7. Exemplary Demonstration of a Difference-in-Differences Approach Utilizing a Natural Experiment 20

Figure 8. Overview of Research Area I 25

Figure 9. Similarity Cube 29

Figure 10. Pipeline for Similarity Calculation of Document Pairs 30

Figure 11. Experimental Design 35

Figure 12. Accuracy of Similarity Estimations Based on Various Document Representations 38

Figure 13. Document Similarity and Industry Distance 40

Figure 14. Accuracy of Different Document Representations with Respect to Object Similarity 41

Figure 15. Document Similarity and Temporal Distance 42

Figure 16. Accuracy of Different Document Representations with Respect to the Temporal Similarity 42

Figure 17. Overview of Research Area II 46

Figure 18. Design Requirements, Principles, and Features 55

Figure 19. Process Map for Prototype Development and Evaluation 59

Figure 20. ROC and Precision-Recall Curves 60

Figure 21. Word Cloud from the Full Corpus and the Extracted Sentences 61

Figure 22. Changes in Analyst Following Over Time, Derived From I/B/E/S and Textual Analyst Reports 79

Figure 23. Changes in Sentiment Polarity Over Time 81

Figure 24. Changes in Novelty of Analyst Reports Over Time 83

Figure 25. Changes in the Dissimilarity of Analyst Reports Issued During the Three Days Surrounding the Earnings Announcement 85

Figure 26. Distribution of the Propensity Score 102

Figure 27. CDO Presence and Digital Transformation Activities Within S&P 500 Over Time 133

Figure 28. CDO Ratio and Digital Ratio Across Industries and Communication Tools 134

Figure 29. Digital Transformation Activities Around CDO Appointments 136

List of Tables

Table 1. Overview of Studies ... 6

Table 2. Dataset Description ... 21

Table 3. Sample Excerpt From the I/B/E/S Database ... 22

Table 4. Document Representation in Research of the Finance and Accounting Domain ... 32

Table 5. Configuration of the Binary Experiment for Similarity Detection ... 36

Table 6. Configuration of Document Representation ... 37

Table 7. Top Companies and Brokers within the Sample ... 58

Table 8. Top Words of Topics from Extracted Sentences ... 62

Table 9. Per-Topic Examples of Environmental Sentences ... 63

Table 10. Contingency Table Between Broker and Firm Location ... 78

Table 11. Full Sample Summary Statistics ... 78

Table 12. Regression Results – Analyst Coverage Based on I/B/E/S Data and Textual Analyst Reports ... 80

Table 13. Regression Results – Changes in Analyst Optimism ... 82

Table 14. Regression Results – Novelty of Textual Analyst Reports ... 84

Table 15. Regression Results – Dissimilarity of Analyst Reports Issued During the Three Days Surrounding the Earnings Announcement ... 86

Table 16. Placebo Testing ... 87

Table 17. Variable Definition ... 92

Table 18. Sample Construction ... 106

Table 19. Descriptive Statistics ... 107

Table 20. Regression Results – Analyst Coverage and MiFID II ... 108

Table 21. Regression Results – Loss of Analyst and Media Coverage ... 110

Table 22. Regression Results – Media Coverage and MiFID II ... 111

Table 23. Regression Results – Media Coverage and MiFID II During Earnings Announcements ... 113

Table 24. Regression Results – Reduced Variation in Analyst Coverage Changes ... 114

Table 25. Placebo Testing ... 115

Table 26. Variable Definition 118
Table 27. Descriptive Statistics 135
Table 28. Correlation Matrix 135
Table 29. Panel Regression 137

Abbreviations

AP .... Average Precision
ASCII .... American Standard Code for Information Interchange
AUC .... Area Under the Curve
BERT .... Bidirectional Encoder Representations from Transformers
CC .... Conference Call
CDO .... Chief Digital Officer
CEO .... Chief Executive Officer
CFA .... Chartered Financial Analyst
CIO .... Chief Information Officer
CO_2 .... Carbon Dioxide
CSR .... Corporate Social Responsibility
CxO .... C-Level Position
DF .... Design Feature
DP .... Design Principle
DR .... Design Requirement
DSR .... Design Science Research
EA .... Earnings Announcement
EDGAR .... Electronic Data Gathering, Analysis, and Retrieval
EEA .... European Economic Area
EPS .... Earnings Per Share
ESG .... Environmental, Social, and Corporate Governance
EU .... European Union
GDP .... Gross Domestic Product
GloVe .... Global Vectors for Word Representation
GRI .... Global Reporting Initiative
H .... Hypothesis
I/B/E/S .... Institutional Brokers' Estimate System
IFRS .... International Financial Reporting Standards
IPO .... Initial Public Offering
IT .... Information Technology
LDA .... Latent Dirichlet Allocation
LSA .... Latent Semantic Analysis
LSI .... Latent Semantic Indexing
M&A .... Mergers & Acquisitions
MD&A .... Management Discussion and Analysis
MiFID .... Markets in Financial Instruments Directive
MTB .... Market to Book Ratio
NASD .... National Association of Securities Dealers
NEA .... Non-Earnings Announcement
NLP .... Natural Language Processing

NYSE .... New York Stock Exchange
OLS .... Ordinary Least Squares
PDF .... Portable Document Format
PPE .... Property, Plant, and Equipment
PSM .... Propensity Score Matching
Q&A .... Questions & Answers
ROA .... Return on Assets
ROC .... Receiving Operator Characteristics
RQ .... Research Question
S&P .... Standard & Poor’s
SDC .... Securities Data Company
SDGs .... Sustainable Development Goals
SEC .... Securities and Exchange Commission
SMACIT .... Social, Mobile, Analytics, Cloud, and Internet of Things
SMEs .... Small and Medium-sized Enterprises
SRI .... Socially Responsible Investing
SVM .... Support Vector Machine
TDM .... Term Document Matrix
TF .... Term Frequency
TF-IDF .... Term Frequency–Inverse Document Frequency
TRBC .... The Refinitiv Business Classification
US .... United States
USD .... United States Dollar
USE .... Universal Sentence Encoder
VW .... Volkswagen

Symbols

COS(X)..Cosine of X
i .... Company index
j .... Report index
k .... Broker index
ln(X) .... Natural logarithm of X
N .... Number of observations
N_{neg} .... Number of negative words
N_{pos} .... Number of positive words
q .... Quarterly index
t .... Time index
z(X)... Z-transformed time series of X
α .... Regression coefficient
Δ .... Difference
ε .... Error term

A. Foundations

1 Introduction

1.1 Motivation

The relationship between management and investors is often characterized by information asymmetries. Typical problems that arise due to information asymmetries are moral hazard and adverse selection. Information asymmetries can lead to economic welfare losses, as they increase the cost of external financing, and management may hence be unable to make profitable investments (Myers and Majluf 1984). Information asymmetries can thus pose a problem from an overall economic perspective. Different circumstances might affect the information asymmetries between investors and companies. These include data processing capabilities, regulatory changes, and altered information needs (e.g., sustainability information). Organizational changes to companies, such as digital transformation, can also lead to additional informational needs of investors. Various strategies are conceivable for reducing asymmetries and increasing overall economic welfare. Three generic strategies for the reduction of information asymmetries are addressed in this thesis and illustrated in Figure 1.

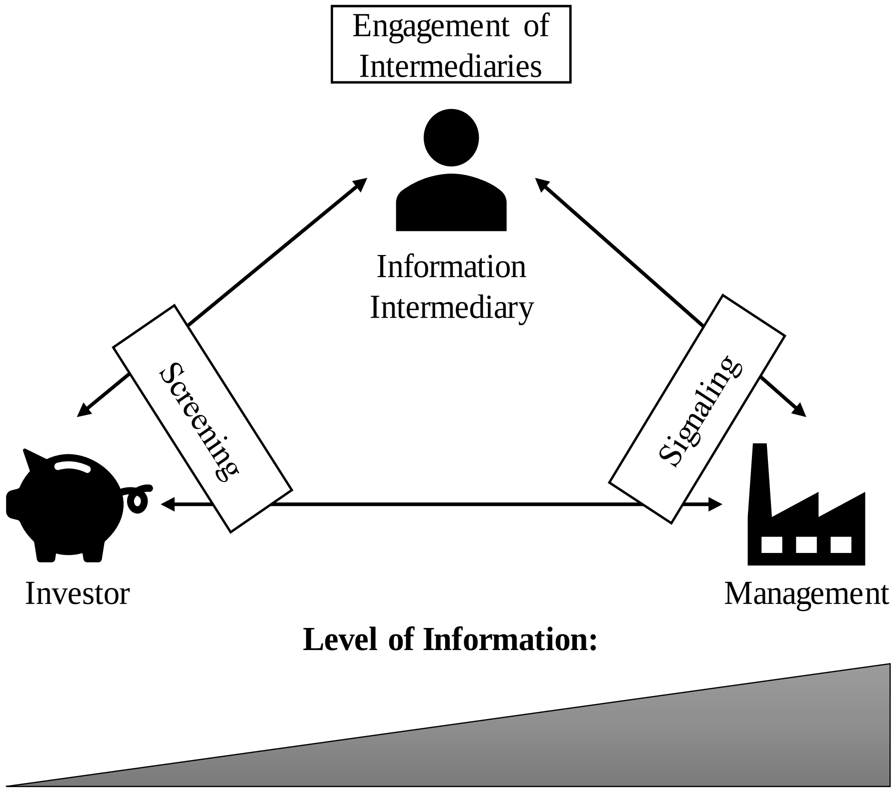

Figure 1. Strategies for Reducing Information Asymmetries

First, investors might execute independent analyses and identify companies with positive characteristics through a screening process and invest accordingly. State-of-the-art methods make it easy to analyze large volumes of data automatically. Since a growing part of the data relevant for investors is not perfectly structured but is available as unstructured (text) data, natural language processing (NLP) methods are particularly important (Lewis and Young 2019). During their decision-making process, investors and other stakeholders must now screen not only companies' financial characteristics but also their sustainability initiatives. Sustainability has become an essential factor for investors' decision-making (Unruh et al. 2016; Hartzmark and Sussman 2019). Therefore, it is highly relevant to evaluate how investors can reduce information asymmetries concerning companies' sustainability performance by leveraging unstructured data with suitable information systems.

Second, investors might not perform the screening by themselves but rely on information intermediaries for this task. Sell-side analysts are an important group in this respect. For small investors without access to analysts, financial journalists can also play an important role in information gathering. With the introduction of the Markets in Financial Instruments Directive (MiFID) II in 2018, the European Union (EU) has implemented a regulation designed to increase transparency in the capital market. However, this regulation has also altered the business model of sell-side analysts. Examining the impact of this regulation on information intermediaries' engagement is therefore of particular relevance for understanding the extent to which information intermediaries are restricted or supported by the regulation in fulfilling their function.

Finally, the company itself might have an incentive to reduce or remove information asymmetries. This behavior is referred to as signaling (Spence 1973). Companies with positive characteristics have an incentive to communicate these characteristics and thus stand out from companies with poor characteristics. Communication with stakeholders is crucial in times of major organizational change (e.g., Schweiger and Denisi 1991). Digital transformation represents such an organizational change process. This raises the question of how the corresponding communication can be implemented at the organizational level. One possibility could be to appoint a chief digital officer (CDO), who is explicitly responsible for digitization initiatives, in the top management team. It is therefore relevant to evaluate the extent to which CDOs play a role not only in pursuing but also in communicating and thus in signaling digital transformation.

1.2 Research Questions

Document similarity is a useful measure to directly (not via capital market reactions) evaluate the informativeness of financial documents (Hanley and Hoberg 2010) or to investigate information flows. Thus, document similarity could be applied as a direct and text-based measure in research concerned with information asymmetries. **Research Area I** lays the methodological foundation for Bankamp et al. (2022, paper II.2) by systematically comparing and analyzing different document representations for similarity calculations. Unlike in sentiment analysis, where the Loughran and McDonald (2011) dictionary has been established for finance and accounting literature (Kearney and Liu 2014), document similarity lacks best practices. First, one must evaluate which dimensions of similarity are present in research on finance and accounting. Based on these dimensions, **Research Question I.1** aims to identify the most suitable representation for similarity analysis. Answering this question should provide methodological guidance for future research.

> **Research Question I.1:** How should one choose document representations based on the dimension of similarity to capture?

Research Area II forms the empirical foundation of the thesis and considers the problem of information asymmetry between companies' management and (potential) investors from different perspectives. **Research Question II.1** focuses on a technical approach to allow more accessible screening for investors with a preference for sustainable stocks.

> **Research Question II.1:** How should a system be designed for extracting sustainability-relevant information from analyst reports?

Investors also rely on sell-side analysts and journalists to reduce information asymmetries regarding their investments. **Research Questions II.2** and **II.3** handle the effect of the MiFID II regulation on the engagement of information intermediaries. The concept of document similarity, which is extensively examined in **Research Area I**, is applied to answer **Research Question II.2**.

> **Research Question II.2:** How do analysts modify their reports to respond to the changed market conditions induced by MiFID II?

> **Research Question II.3**: How does MiFID II's research unbundling impact the provision of information by the media?

With signaling, companies themselves can also reduce information asymmetries. Signaling is reasonable for companies if they possess positive characteristics to disclose. Thus, they can differentiate themselves from others. Against this backdrop, **Research Questions II.4a** and **II.4b** examine the CDO's role when companies engage in topic-specific signaling regarding their digital transformation.

> **Research Question II.4a**: How does CDO presence impact the volume of digital transformation-related signals in external communication tools?

> **Research Question II.4b**: How does the volume of digital transformation-related signals differ across communication tools with different degrees of regulation?

2 Structure of the Thesis

The thesis is divided into three sections. Section A (Foundation) builds the foundation of the thesis by addressing its motivation, research questions, structure, and research background. Section B (Studies) is the main part of the thesis; it contains the five individual research studies. The five studies are divided into two research areas. Bankamp and Muntermann (2022, paper I.1) build the first research area and deals with the methodological concept of document similarity. The remaining four papers are assigned to Research Area II, which focuses on information asymmetries in capital markets and how they can be reduced through technological deployment (Bankamp and Muntermann 2021, paper II.1) as well as information intermediaries, the role of regulation (Bankamp 2022, paper II.2; Bankamp et al. 2022, paper II.3), and management's signaling (Metzler et al. 2021, paper II.4). The overall structure of the thesis as well as the interrelationships among the individual studies are depicted in Figure 2.

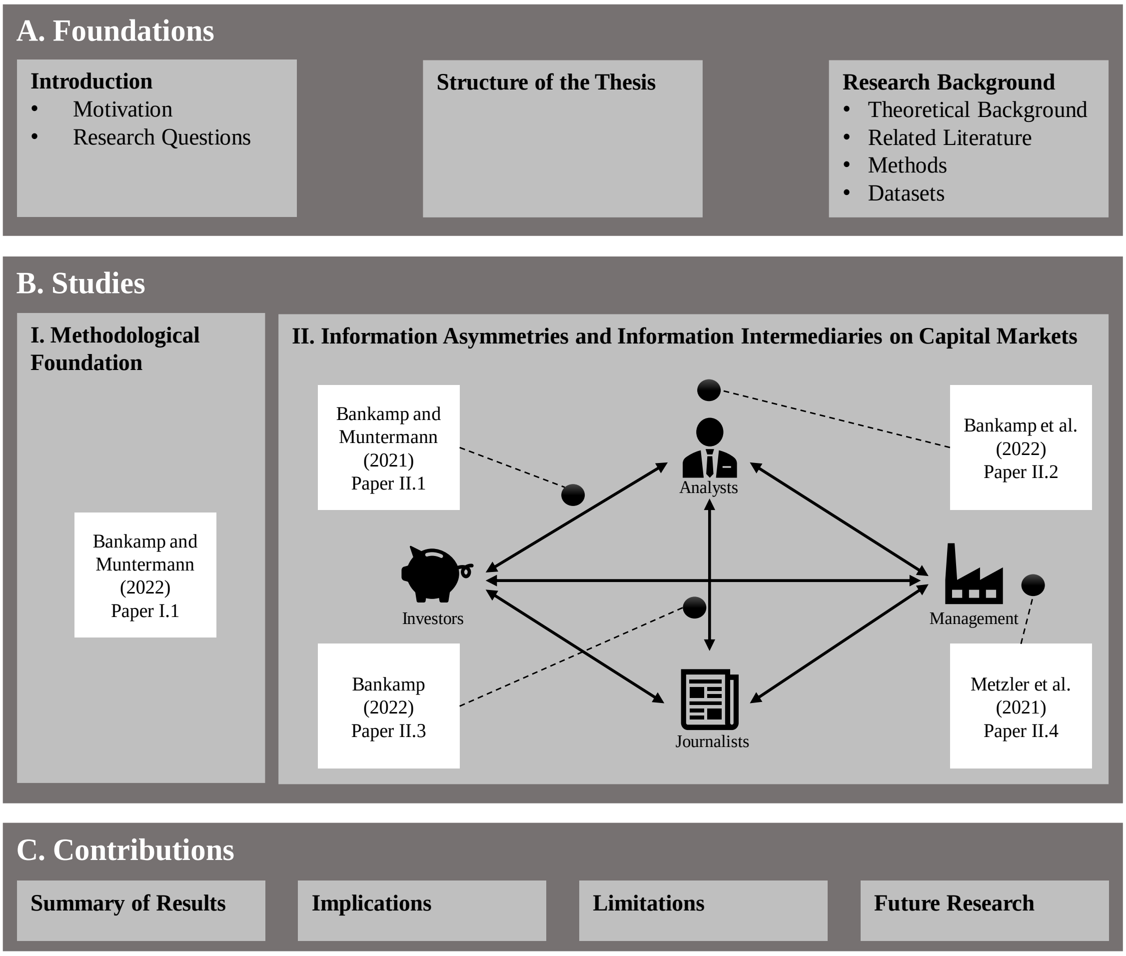

Figure 2. Overview of the Thesis

Table 1 provides an overview of the five studies in this thesis and lists their main contribution. Three of these papers have already been published. The other two papers were presented (in a previous version) at international research conferences. For this thesis, the layout of the studies and labeling of variables is harmonized across all papers to improve the reading flow. Section C (Contributions), which forms the third part of the paper, summarizes and integrates the findings of the studies. In particular, implications and limitations are discussed. In addition, an outlook for future research is presented.

Study	Conference	Status	Contribution
I. Methodological Foundation			
Bankamp and Muntermann (2022) Paper I.1	PACIS 2022	Published	Development of an evaluation framework for text similarity assessments and problem-related identification of suitable document representations for text similarity.
II. Information Asymmetries and Information Intermediaries on Capital Markets			
Bankamp and Muntermann (2021) Paper II.1	PACIS 2021	Published	Development of an artifact and design principles for extracting sustainability information from unstructured analyst reports.
Bankamp et al. (2022) Paper II.2	INQUIRE Autumn Residential Seminar	Presented	Investigating the impact of MiFID II on the textual content of analyst reports.
Bankamp (2022) Paper II.3	12th Annual Pre-ICIS Workshop on Accounting Information Systems	Presented	Assessing the impact of MiFID II on the output of financial journalists.
	44th Annual Congress of the European Accounting Association	Presented	
Metzler et al. (2021) Paper II.4	ICIS 2021	Published	Investigating the impact of CDO presence on digital transformation-related signals in firms' external communication.

Table 1. Overview of Studies

3 Research Background

3.1 Theoretical Background

3.1.1 Information Asymmetries

Neoclassical theories are based on the idealized assumption of perfectly informed market participants (Dyner and Franco 2004). In reality, however, this assumption is rarely fulfilled. Prior to a transaction, the seller might have more precise information than the buyer regarding the product being sold (Akerlof 1970). In this case, information asymmetries exist between the two market participants and result in problems such as adverse selection or moral hazard. As these problems can be reduced by institutions, the theory of information asymmetries also provides an economic justification for institutions within the real-world economy (Akerlof 1970). Figure 3 overviews contractual relationships with information asymmetries, their consequences, and their potential mitigations.

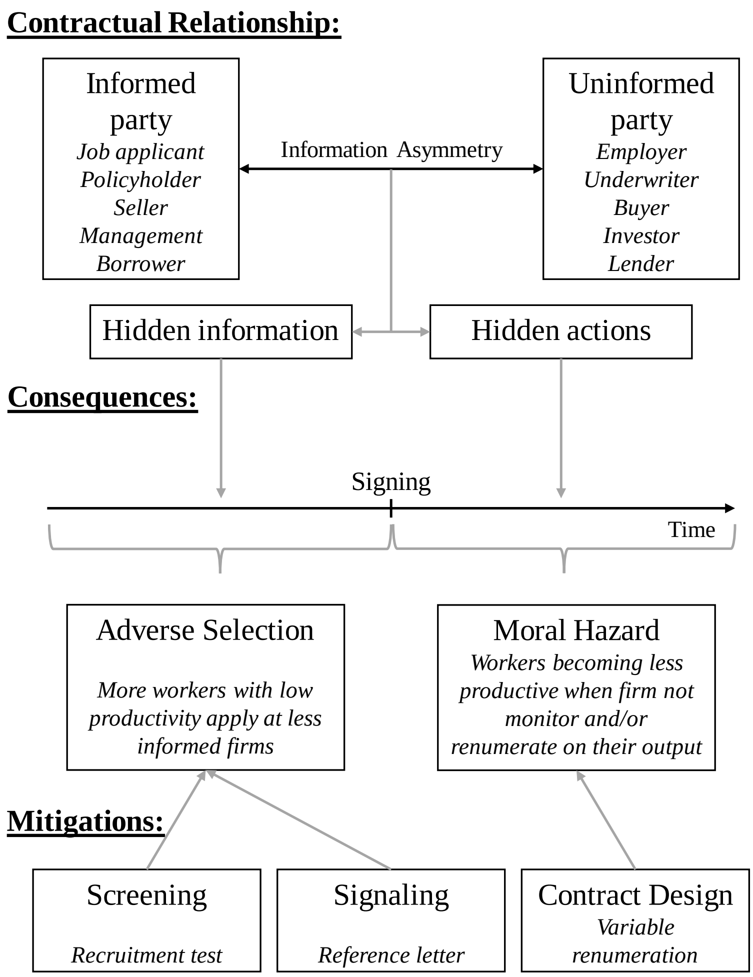

Figure 3. Overview of Information Asymmetries, Their Consequences, and Their Mitigations

Adverse selection can occur in many areas. The most famous example might be the market for used cars from Akerlof (1970). The seller of a used car has a clear idea of its condition and therefore sells low-quality vehicles while retaining possession of high-quality cars. The theory is also intensively applied to the insurance market. If a policyholder has private information about their risk, they might buy more (less) insurance if the risk is high (low). A. Cohen and Siegelman (2010) conducted a literature review of various insurance markets and found evidence of adverse selection in some, but not all markets. Policyholders with a higher risk exposure do not always buy more insurance, as policyholders might not always have the capability to assess their risk better than the insurer (A. Cohen 2005). In such a case, there is no information asymmetry, which means that the problem of adverse selection does not exist.

The problem of adverse selection triggered by hidden information can be reduced by screening and signaling. Screening refers to the acquisition of information through the uninformed party (Stiglitz 1975). Screening can also be outsourced and performed by an intermediary, such as a financial intermediary (Ramakrishnan and Thakor 1984) or an information intermediary (Healy and Palepu 2001). Signaling refers to the provision of information (signals) by the informed party that allows the uninformed party to assess the properties of the informed party and thus reduces information asymmetries (Spence 1973). Signaling is described in detail in Section 3.1.3. Furthermore, moral hazard can be reduced by adapting the contractual design (Baron and Besanko 1987). Increased deductibles of insurances, for example, can provide incentives for risk avoidance and thus reduce moral hazard (Wang et al. 2008).

In terms of screening, Bankamp and Muntermann (2021, paper II.1) have built a system that investors (uninformed parties) can use to leverage sustainability information about possible investment objects. Bankamp et al. (2022, paper II.2) and Bankamp (2022, paper II.3) are focusing on outsourced screening conducted by information intermediaries (financial analysts and journalists), while Metzler et al. (2021, paper II.4) are based on signaling theory.

3.1.2 Information Intermediaries

This thesis follows Rose (1999), who describes information intermediaries as "economic agents supporting the production, exchange, and utilization of information in order to increase the value of the information for its end-user or to reduce the costs of information acquisition" (Rose 1999, p. 76).

In a perfect market, there is no economic explanation for the existence of (information) intermediaries, since the individual market participants can perform the tasks themselves (Campbell and Kracaw 1980). However, real-world markets often do not match their idealized model. Information asymmetries often exist between (potential) contracting parties. This is where information intermediaries can provide an economic benefit. By collecting, processing, and distributing information, the information intermediary can reduce asymmetries between the two parties and thus contribute to an efficient allocation of resources (Bhargava and Choudhary 2004). In contrast to transaction intermediaries, information intermediaries are not involved in the transaction process that occurs between the two parties but provide (general) information to these parties (Eckardt 2007). It can be argued that the contracting parties themselves could also reduce these information asymmetries (e.g., through screening). However, according to Rose (1999) definition, an information intermediary can lead to cost savings in information acquisition. Such cost savings can originate from the information intermediary’s scaling effects (Womack 2002). Rating agencies have large economies of scale, as not every investor who wants to invest in a bond has to analyze the creditworthiness of all possible companies. Instead, the task of credit assessment is bundled within the rating agencies, and the information is later distributed to the investors (Rhee 2015). However, it should be considered that information intermediaries are agents themselves and thus add further principal–agent conflicts (Spindler and Jaenig 2010). The structure of the initial principal–agent relationship thus becomes increasingly complex. Rating agencies might have the incentive to publish inflated ratings (underestimation of the probability of default) instead of an unbiased rating, as it would be in the interest of investors (Stolper 2009). Information intermediaries can exist in a variety of institutional forms. Womack (2002) distinguishes between information intermediaries that operate as for-profit, non-profit, or governmental organizations. The socially optimal institutional setup depends on the various characteristics of the information market (Womack 2002). This thesis focuses on information intermediaries (sell-side financial analysts and journalists), which are usually organized as for-profit organizations.

Information intermediaries can play a critical role in many fields or industries where asymmetries between contractual parties are present, such as in healthcare (e.g., Zahedi and Song 2008) and the retail sector (e.g., Viswanathan et al. 2007). Rating agencies, financial analysts, and media outlets are among the most important capital market-related information intermediaries. Rating agencies bring new information to the market regarding the creditworthiness of borrowers, thus reducing overall information asymmetry (Cavallo et al. 2013; Ingram et al. 1983). This applies not only to the rating score itself but also to the textual information provided in rating agencies’ reports (Agarwal et al. 2016). The same applies to financial analysts, where informational value has been found in their earnings forecasts (Lys and Sohn 1990), price targets (Brav and Lehavy 2003), and textual explanations (A.H. Huang et al. 2014a). However, not only financial experts such as rating agencies or financial analysts can reduce information asymmetries, but also the media can reduce asymmetries between a company's management and its investors (OuYang et al. 2017).

3.1.3 Signaling Theory

Signaling theory was developed by Spence (1973) and was first applied to the problem of recruiting in the job market when an employer has incomplete information regarding candidates' capabilities. The theory was later brought to many different fields in economics and beyond (Connelly et al. 2011). The basis of the theory is that quality differences exist among the informed parties, but the uninformed party has no knowledge about the individual qualities. The informed parties, by contrast, know about their qualities. Spence (1973) models this problem by dividing the group of informed parties into two subsets (high-quality and low-quality). In addition, it is assumed that the benefit from a transaction with the high-quality group is greater than the benefit from the low-quality group. Accordingly, the uninformed party would also be willing to pay a higher price to the high-quality group. However, since the uninformed party cannot separate the two groups, the result is a uniform market price that lies between the willingness to pay for the low and high quality. This in turn leads to an incentive for the informed party to disclose its own quality if it belongs to the high-quality group. In this way, they would receive the uninformed party's actual willingness to pay for high quality and not just the mixed calculation as compensation. This disclosure of quality is called signaling. Since anyone could claim to belong to the high-quality group, the signal must be reliable. This means that the signal actually allows a statement on the quality and is not easily imitable by the low-quality group (Connelly et al. 2011).

One of the best-known examples of signaling relates to education. Workers can send signals of education to the job market, thereby reducing information asymmetries about their skills. The purpose of education in this model is not to provide job-related skills, but to be an indicator of what quality the applicant brings to the job (Spence 1973). Numerous empirical studies have been conducted to examine the purpose of education for signaling in the job market. The results draw a highly heterogeneous picture about the existence of education's signaling function (see Heywood and Wei 2004). Heywood and Wei (2004) argue that signaling predominantly plays a role in unregulated labor markets, and they show that signaling does indeed play a role for the weakly regulated labor market of Hong Kong. However, they also indicate that education serves not only as a form of signaling but also as a means to increase workers' productivity.

Research has demonstrated that consumers respond to signals (e.g., product warranties) from producers (Boulding and Kirmani 1993). Online community platforms can implement information technology (IT) features as signals to increase users' trust and participation (Benlian and Hess 2011). Management might apply stock splits to signal their private information regarding future earnings to investors (McNichols and Dravid 1990). Overall, there exists a major research area applying signaling theory to relationships between a company's management and the capital market (see Connelly et al. 2011). This research area is discussed in more detail in Section 3.2.3.

3.2 Related Literature

3.2.1 Natural Language Processing of Financial Documents

In recent years, the volume of unstructured text data in corporate disclosure has significantly increased (Lewis and Young 2019). To make this information easily accessible, the use of NLP is becoming increasingly important for researchers and practitioners in this domain. Two basic research strands have emerged that apply NLP to financial documents. The first strand of research belongs to the classical finance literature and uses methods from NLP to operationalize a construct. NLP methods are a tool to prepare the (unstructured) text for empirical analysis. Typically, the goal is to construct a variable that is later incorporated into the statistical model (e.g., a linear regression), equivalent to structured data (e.g., price or balance sheet ratios). R. Yang et al. (2018), for example, derive firms' risk disclosure from financial statements using NLP to assess the association between risk disclosure and audit fees. However, scholars who apply NLP in the finance and accounting domain primarily focus on sentiment analysis, where the tone of texts (positive or negative) is analyzed (Loughran and McDonald 2016). Application ranges from earnings releases (X. Huang et al. 2014b) to conference calls (CC) (De Amicis et al. 2021), financial reports (Feldman et al. 2010), and analyst reports (A.H. Huang et al. 2014a), to name a few.

The other research strand is design science research (DSR). Here, the focus is not on positivist research on the behavior of market participants in the capital market but on generating knowledge about how certain systems should be designed to solve market participants' or other stakeholders' problems in the capital market. The area of application within the finance domain is broad. It includes text-based fraud detection in securities trading (Siering et al. 2021) or the detection of fraudulent corporate reporting (Dong et al. 2016). Chou et al. (2018) developed a system to assess the consistency between quantitative and narrative disclosure in financial documents. Although not framed as design science, Loughran and McDonald (2011) famous paper can also be regarded as such. Aside from the positivist finding that classical sentiment dictionaries are unsuitable for the analysis of financial documents, a major contribution of Loughran and McDonald (2011) lies in the development of their word lists, which many scholars utilize when conducting sentiment analysis in the financial domain (Kearney and Liu 2014).

It should be noted that these two strands cannot be completely separated. For example, design elements can also be recognized in some positivist studies without being explicitly mentioned. While L. Cohen et al. (2020) designed a trading strategy to prove that the capital market is not capturing all information provided by adjustments in financial reports' structure and language, the authors implicitly also created design knowledge, as the generated design knowledge (trading strategy) enables market participants to invest accordingly and to take advantage of this inefficiency. For a comprehensive overview of the use of NLP in the finance domain, the literature reviews by Fisher et al. (2016) and Gupta et al. (2020) are recommended.

Bankamp and Muntermann (2022, paper I.1) and Bankamp and Muntermann (2021, paper II.1) belong to the second research strand, as they create design knowledge. Bankamp and Muntermann (2022, paper I.1) deal with methodological knowledge, while Bankamp and Muntermann (2021, paper II.1) are specifically concerned with the design of an artifact and explicitly frame their work as design science research. Bankamp et al. (2022, paper II.2) and Metzler et al. (2021, paper II.4) belong to the first research strand discussed in this section (positivist application of NLP in the finance domain). Bankamp et al. (2022, paper II.2) use a sentiment score based on machine learning as a construct for analysts' optimism and cosine distance as a construct for the novelty of analyst reports. Metzler et al. (2021, paper II.4) apply a dictionary to extract specific sentences from text (information retrieval) to derive a construct for the volume of digital transformation-related signaling.

3.2.2 Information Intermediaries and Regulatory Change

As information intermediaries are agents themselves, they do not come without principal–agent conflicts (Spindler and Jaenig 2010). These conflicts put information intermediaries on the agenda of regulators. Although financial journalists are affected by general regulations such as the prohibition of insider trading or market manipulation (Tambini 2010), recent regulatory reforms have especially affected financial analysts. For this reason, this section focuses on financial analysts.

A major point of discussion in regulatory measures has been the revenue model of analysts, as it can lead to conflicts of interest and misleading incentives. At the heart of this debate is the question of whether analyst reports are sold as a bundle with execution services (soft dollar commission) or whether the reports are priced as a separate service (hard dollar). Due to the price fixing of trading commissions in the United States (US) until 1975, bundling became common practice as a way for brokers to differentiate themselves from competitors (Jorden 1975). Even after the abolition of this price fixing, the practice of bundling continued. This was made possible by the Securities and Exchange Commission (SEC), which provided asset managers with a safe harbor that allowed them to look beyond the price criteria when selecting brokers

(Johnsen 2008). In the European Economic Area (EEA), the introduction of the MiFID II, that consist of a large set of financial regulations, on January 3, 2018 has prohibited this practice (EU 2014). Brokers must now charge asset managers separately for execution and research services. Bender et al. (2021) provide a comprehensive literature review on research unbundling. There is mixed evidence on the effectiveness of research unbundling (separating research from trading commissions). While it was found that research quality has increased, the quantity of sell-side analyst research has decreased (e.g., B. Fang et al. 2020). Evidence also suggests that mutual fund investors do not profit from the new payment regime introduced by MiFID II, as it does not offer the increased transparency for which it was designed (Fröberg and Halling 2021). The impact of MiFID II on the quality of analysts' (textual) reports is assessed in Bankamp et al. (2022, paper II.2). Bankamp (2022, paper II.3) analyzes the impact of MiFID II on the quantity of financial journalists' output.

Another important regulation affecting analysts is Regulation Fair Disclosure, which is intended to create a level playing field between institutional and private investors by prohibiting companies from selectively disseminating information (e.g., to individual analysts) (Bailey et al. 2003). Regulation Fair Disclosure was found to lead to a decrease in accuracy of earnings forecast (Agrawal et al. 2006; Bailey et al. 2003); in addition, it reduced analysts' optimistic bias, as analysts are no longer dependent on management's exclusive information and thus have less incentive to please management with optimistic forecasts (Herrmann et al. 2008).

As the Sarbanes–Oxley Act required the regulation of analysts, National Association of Securities Dealers (NASD) Rule 2711 and New York Stock Exchange (NYSE) Rule 472 were introduced in the US in 2002 to address the interdependencies between the investment banking business and the research business within brokerage houses and to mitigate conflicts of interest (Espahbodi et al. 2015; Bradshaw et al. 2017). Analysts working for a brokerage house that also provides investment banking for the covered companies have an incentive to publish optimistic analyses. Dugar and Nathan (1995) provide evidence that analysts who have such a relationship with the covered company tend to issue more positive forecasts and recommendations. NASD Rule 2711 and NYSE Rule 472 decouple analysts' compensation from brokerage houses' investment banking business, which is intended to reduce conflicts of interest. This attempt has been successful, as analysts' recommendations have become more independent from investment banking opportunities than before (C.Y. Chen and Chen 2009). The two aforementioned regulations also required analysts to disclose the distribution of sell, hold, and buy recommendations. With NASD Rule 2711 and NYSE Rule 472 coming into effect, buy recommendations strongly decreased, while sell recommendations increased (Barber et al. 2006).

In addition to regulations directly targeting sell-side analysts, a large body of literature has examined how certain accounting standards affect analysts' forecast accuracy. It has been found, for example, that mandatory adoption of the International Financial Reporting Standards (IFRS) leads to increased accuracy among foreign analysts (H. Tan et al. 2011) or that strict enforcement of accounting standards increases analysts' accuracy (Hope 2003). However, the focus of these studies has not been on the analysts themselves. Instead, forecasts have been used as an indicator of the usefulness of certain accounting standards for decision-making.

3.2.3 Management's Signaling Toward Capital Markets

As discussed in Section 3.1.3, an agent can use signaling to reduce information asymmetries with respect to the principal. A company's management act as an agent with an information advantage, while the investor acts as a principal. This can lead to problems of adverse selection and moral hazard. The literature has demonstrated numerous ways in which management applies signaling.

Management might use dividends to signal positive private information regarding the firm's expected cash flows (Bhattacharya 1979; M.H. Miller and Rock 1985). However, empirical studies have questioned the dividend-signaling hypothesis (Gunasekarage and Power 2002; Brav et al. 2005). Empirical research has shown that the chief executive officer (CEO) can signal higher credibility by holding more shares and a high number of external directorships to lower information asymmetries (Yan Zhang and Wiersema 2009). Firms might also apply corporate social responsibility (CSR) practices to signal the firm's quality and capabilities (Su et al. 2016). Moreover, on the one hand, management might have incentives not to send certain signals. The packing-order theory states that management should prefer internal financing and debt before sourcing capital via the issue of new equity, as this would signal overvalued stocks to the capital market, which in turn would lead to falling share prices that would harm current investors (Myers and Majluf 1984). On the other hand, when repurchasing stocks, management can signal positive private information toward investors (Vermaelen 1981).

In addition, the management board itself can serve as a signal. This is especially the case when information asymmetries exist between current and potential investors. Existing shareholders can send certain signals to potential investors through the composition of the top management team. Certo (2003) argues that a prestigious board signals the high quality of the company, as prestigious managers would retreat from taking up such a position in a low-quality company because of reputational risks. B.D. Cohen and Dean (2005) empirically test this construct in an initial public offering (IPO) scenario and find that top management team's legitimacy (measures via industry experience, management experience, age, and education) can serve as a signal to reduce information asymmetry in the run-up to an IPO. Thus, high-quality companies can signal their value by appointing a legitimate top management team to reduce IPO underpricing (B.D. Cohen and Dean 2005). Furthermore, the heterogeneity (functional and educational background) of the top management team serves as a quality signal and reduces IPO underpricing (Zimmerman 2008). Metzler et al. (2021, paper II.4) discuss the role of the CDO for signaling of topic-specific information regarding digital transformation.

3.3 Methods

3.3.1 Natural Language Processing

In four of the five studies in this thesis, unstructured text data is analyzed to answer the research questions. Only Bankamp (2022, paper II.3) does not utilize unstructured data. Therefore, text mining methods are an important methodological part of the present thesis. In the following sections, the individual methods that are applied are described and discussed briefly.

Pre-Processing

A wide variety of methods exist in the field of NLP. However, most analyses should not be applied directly to raw texts. Thoroughly applied pre-processing of the raw text can lead to substantially more accurate results (Jianqiang and Xiaolin 2017). However, pre-processing itself is a very complex field, and there is no "one size fits all" approach, as the appropriate choice of text pre-processing methods depends on the dataset and upstream analysis (Naseem et al. 2020). The most common pre-processing steps applied in the studies of this thesis are presented in Figure 4 and discussed below.

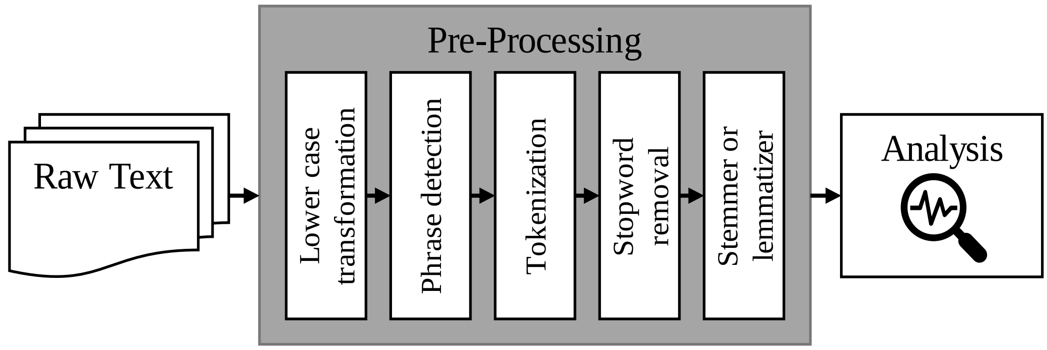

Figure 4. Text Pre-Processing

A typical first step in pre-processing is the transformation to lower case. This detection leads to consistent recognition of words regardless of the case (Gull et al. 2016; Naseem et al. 2020). Otherwise, a word that is capitalized because it marks the beginning of a sentence could not be mapped to the equivalent word in lower case. Thereafter, a phrase detection can be performed. This allows for the concatenation of words that form a logical entity. Phrase detection might transform "balance sheet" to "balance_sheet." The advantage of applying phrase detection is that phrases whose meaning differs from that of the individual words they consist of can be represented correctly (Mikolov et al. 2013b). Such phrase detection should be performed at the beginning of pre-processing, when punctuation is still present, to avoid incorrectly detecting cooccurrences that span over two sentences.

Since words are typically analyzed in NLP, the next step is to remove punctuation, numbers, and other symbols. The text is then tokenized (Webster and Kit 1992), which means that the text is cut in smaller logical entities (tokens). Usually the tokenization is done on a word level, so that each word is forming a single token. However, there are also tokenizers for the sentence level or the n-gram (multiple word) level. The tokenization step shows the benefit of phrase detection. Through concatenation, words that have been detected as a phrase are not separated during tokenization and remain as a single token.

Texts contain many words, such as articles, that are important to the human reader but do not contribute to the content of the text. Stop word removal removes these common words that do not contribute to the content of the text, thus reducing the text to its essence (Silva and Ribeiro 2003). Subsequently, the complexity of the text is reduced by removing different conjunctions. This can be done through one of two approaches. In stemming (e.g., Porter 1980), word endings are truncated based on an algorithm, thus reducing a word to its stem. Different conjunctions or the plural of nouns are truncated. An alternative to stemming is lemmatization, a technique that utilizes the context of the word to find the correct lemma for each word. In this way, a lemmatizer will be able to transform "goes" → "go" and "went" → "go," while a stemmer is usually only able to transform "goes" → "go" and will fail to find the correct stem of "went" (Jivani 2011).

Text Representation

After pre-processing is complete, the text data is still unstructured. However, most statistical or machine learning methods require a structure to perform calculations. The text representation transforms the unstructured text data into numerical vectors and thus makes them usable for the downstream analyses (Zhao and Mao 2018). Many different approaches are available to convert the text into these vectors. Bankamp and Muntermann (2022, paper I.1) cover this topic in more detail; therefore, only a brief description of text representation is provided here. Term frequency (TF) is a simple but frequently used approach; it is a bag-of-words model, since the sequence of the words occurring in the text is lost during the transformation. The individual vectors of such a representation are typically consolidated in a term-document matrix (TDM). Each row of the matrix is linked to a unique word in the document corpus, while each column is linked to a specific document in the corpus. Thus, each cell within the matrix represents the frequency of occurrence of a specific word within a specific document. The matrix thus shows which word occurs at what frequency in which document. Figure 5 illustrates a TDM and the underlying texts. The vectors from the TDM can be used in numerous models, such as support vector machines or regressions. In addition to the simple case of TF, many other more complex models are discussed in Bankamp and Muntermann (2022, paper I.1).

Unstructured

Doc	Text
1	the bank issued a new bond
2	new record on the stock market
3	bond market is strong today

Structured

Word	Doc1	Doc2	Doc3
bank	1	0	0
bond	1	0	1
market	0	1	1
stock	0	1	0
…	…	…	…
today	0	1	1

Figure 5. Illustration of a Term-Document Matrix

Text Classification

Text classification is an essential class of methods within NLP. A widespread classification task is sentiment analysis, in which the text is classified in terms of the sentiment conveyed. However, a text can be classified with respect to many different dimensions, for example whether a text is factual or fake news (Oshikawa et al. 2020), or which theme the text is about (Lu 2013). Methodically, text classification can be distinguished into two approaches. First, texts can be analyzed with the help of dictionaries, which are lists of words that are assigned to a certain class. For sentiment analysis, these are lists of positive and negative words. The sentiment of a text can then be determined from the ratio of positive and negative words contained in the text. For sentiment analysis, generic (e.g., Stone et al. 1966) as well as domain-specific dictionaries (e.g., Loughran and McDonald 2011) exist. However, dictionaries can be used to classify texts not only in terms of sentiment but also thematically, as applied in Metzler et al. (2021, paper II.4). Second, texts can be classified using machine learning. This requires a labeled dataset that is used for model training; the trained model can then be used to classify unlabeled data. This is also referred to as supervised learning (Kadhim 2019). The advantage of machine learning models is that they are usually more accurate than dictionary-based methods (see e.g., A.H. Huang et al. 2014a). However, dictionary-based analyses have better replicability due to their simplicity. In Metzler et al. (2021, paper II.4), a dictionary is used to classify sentences on digital transformation. Bankamp and Muntermann (2021, paper II.1) combine a dictionary-based approach with a machine-learning-based approach to extract sustainability-relevant sentences. For sentiment classification, Bankamp et al. (2022, paper II.2) apply the machine learning model called FinBERT from Y. Yang et al. (2020).

Topic Modeling

Topic models are another important group of methods within NLP. They belong to the class of unsupervised learning methods, since no labeled training dataset is necessary. The goal of topic models is to recognize topics within a document collection and to assign the documents to a few individual topics. A main hyperparameter of topic models is the number of topics that must be defined by the user (Jacobi et al. 2016). Topic models are similar to cluster analysis that is used on structured data. The difference, however, is that topic models typically do not assign a document to a single topic (Blei et al. 2003). The low number of topics (compared to the high dimensionality of text) makes it possible for a person to obtain an overview of a large document collection containing thousands or millions of documents. A common practice for this is to label each topic based on the most important words within the individual topics (for labeling of topics see e.g., A.H. Huang et al. 2018). Topic models can also be understood as document representation. Therefore, the topic vectors are viewed as a dimension reduction of the TDM and can serve as document representation. In this way, the typically very and sparse TDM is reduced from thousands of features to a few topics. The topic model Latent Dirichlet Allocation (LDA) is applied in Bankamp and Muntermann (2021, paper II.1), while different topic models are used and evaluated as document representation in Bankamp and Muntermann (2022, paper I.1).

Document Similarity

Methods for determining document similarity can be used for numerous applications. The application for search engines or for matching answers to questions is particularly relevant. The method of document similarity can be divided into two steps: in the first step, the text must be converted into a document representation; in the second step, the similarity of two documents is measured by determining the distance between the vectors of the document representation. Many different distance measures can be employed for this, but the cosine distance has become the standard measure in NLP (Loughran and McDonald 2016).

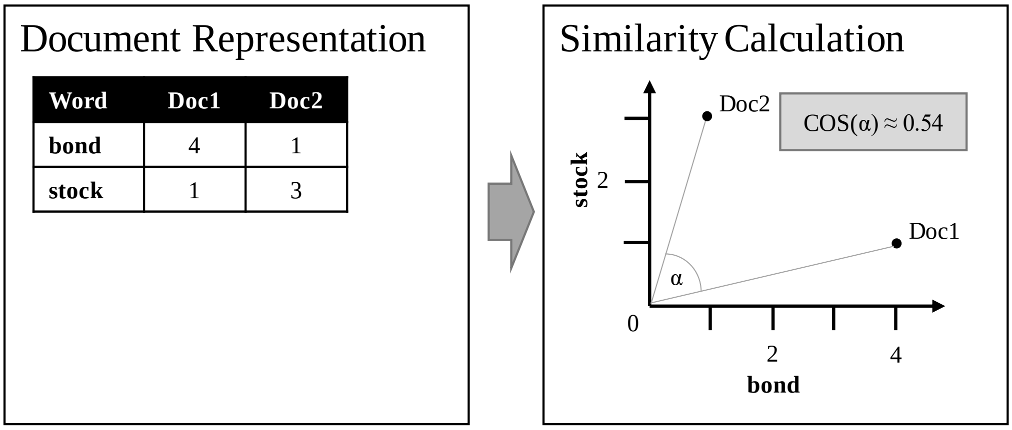

Figure 6. Illustration of Document Similarity Calculations

The calculation of the cosine distance is illustrated in Figure 6. Each document can be represented by a point in the coordinate system. The coordinates are derived from the document representation. In this simple case, the two dimensions are the TF of the words “bond” and “stock.” For distance calculation, the angle between these two points is measured, starting from the coordinate origin. The cosine distance is then formed by calculating the cosine of the measured angle. If one wants to represent a similarity measure, the cosine distance can be subtracted from 1. By considering the angle, only the relative word composition is considered, while the absolute frequency is not taken into account. Bankamp and Muntermann (2022, paper I.1) approach document similarity from a methodological point of view, which is the foundation for Bankamp et al. (2022, paper II.2). In Bankamp et al. (2022, paper II.2), the method of document similarity is used to quantify the novelty of analyst reports.

3.3.2 Difference-in-Differences Analysis

Card and Krueger (1994) worked on the minimum wage, applying the difference-in-differences approach and laying an important foundation for subsequent research evaluating natural experiments using this approach. In business research, randomized experiments are not always possible or feasible. However, to investigate the effect of interest, a natural experiment can be a solution. The prerequisite of such an experiment that is evaluated by the difference-in-differences approach is a shock that affects one part of the population (treatment group), while the other part remains unaffected. Unlike in a real experiment, in a difference-in-differences analysis the researcher does not randomize the assignment of the population. Introductions of laws or other regulations, as in the study of Card and Krueger (1994), that have effects only within a certain jurisdiction can provide an opportunity for such an experiment. To illustrate this analysis, a fictitious example is presented below.

Assuming that a researcher is interested in the extent to which cinemas and movie streaming services are complementary or substitutive services, a first idea could be to survey whether people visit the cinema and/or they have a subscription to a streaming service. However, it quickly becomes apparent that an endogeneity problem exists, since the decision to go to the cinema depends on characteristics of the person (e.g., a general interest in films), and these characteristics in turn affect the probability of subscribing to a movie streaming service. However, more convincing evidence could be obtained from a natural experiment. In the context of the COVID-19 pandemic, restrictions were imposed in Germany and many other countries to contain the pandemic. Within Germany, however, not the same restrictions were implemented in every state. Assuming that cinemas in State A were allowed to remain open, while cinemas in the neighboring State B had to close due to the lockdown, this could be a prime setting for a natural experiment. Due to the regulation, the residents of State B were subjected to a treatment, while State A lacked this treatment; therefore, the residents of State A serve as a control group. Before implementation of the treatment (i.e., before the start of the cinema closures in State A), the distribution of streaming subscriptions in both states would have to be surveyed. The same survey would then be carried out after the cinema closures had been introduced. The effect would then be measured by comparing the difference in the distribution of streaming subscriptions between the two federal states before and after the treatment. Figure 7 illustrates the two time points and groups of this hypothetical research study. In this case, it can be seen that the difference between the two groups has increased with the introduction of the restrictions in State A. Accordingly, this difference-in-differences would be attributed to the treatment (cinema closures).

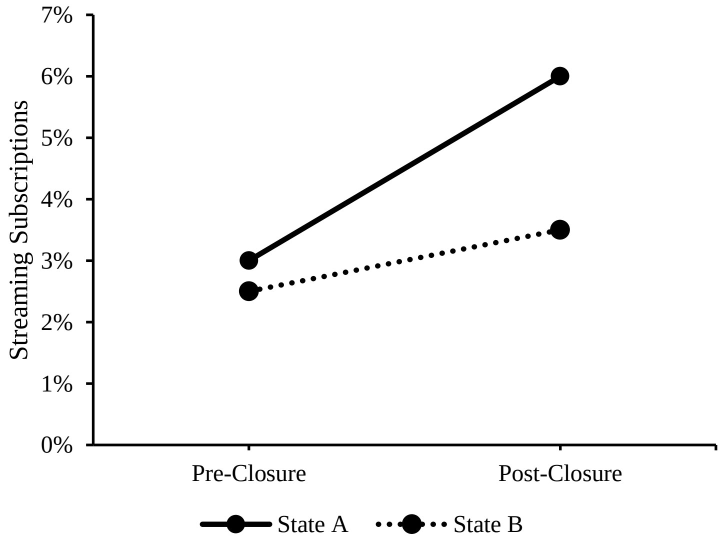

Figure 7. *Exemplary Demonstration of a Difference-in-Differences Approach Utilizing a Natural Experiment*

Using a regression allows one to add fixed effects and further control variables and to easily calculate standard errors of the difference-in-differences estimator (Angrist and Pischke 2008). The standard errors and *t* statistics allow the researcher to judge whether the difference-in-differences is significantly different from zero, and a statement can be made about whether streaming services and cinemas are substitutes, complementary, or neither. The key assumption of this approach is the parallel trend assumption (Angrist and Pischke 2008). This requires the distribution of streaming subscriptions in both federal states to have developed in parallel without the existence of any treatment (cinema closure in A). At this point, a caveat of this setting becomes apparent: if the intervention of cinema closures in A is only part of an overall package of interventions (e.g., sports facilities were also closed in A), the effect could also be attributable to the closure of the sports facilities instead of the closure of the cinemas. The difference-in-differences approach is applied in Bankamp et al. (2022, paper II.2) and Bankamp (2022, paper II.3) to evaluate the effect of the MiFID II regulation on sell-side analysts and journalists.

3.4 Datasets

To answer the research questions posed in this thesis, different datasets are collected. The main datasets utilized within this thesis are listed in Table 2 and discussed below. The datasets can be categorized into three areas: analyst data, media data, and company data.

Datasets	**Source**										**Structure**			**Paper**				
	Thomson ONE	I/B/E/S	EDGAR	RavenPack	Boardex	Amadeus	Crunchbase	Workspace	Eikon	Datastream	Structured	Semi-structured	Unstructured	Bankamp and Muntermann (2022, paper I.1)	Bankamp and Muntermann (2021, paper II.1)	Bankamp et al. (2022, paper II.2)	Bankamp (2022, paper II.3)	Metzler et al. (2021, paper II.4)
Analyst reports	x												x	x	x	x		
Analyst forecasts		x									x				x	x	x	
Media				x							x						x	
Conference calls	x											x						x
10-K reports			x										x					x
Board data					x	x	x				x							x
Constituent, price and accounting data								x	x	x	x			x	x	x	x	x

Table 2. Dataset Description

The analyst data consists of analyst reports, which are retrieved in Portable Document Format (PDF) from Refinitiv Thomson ONE. These text documents are further processed as part of the pre-processing (see e.g., Appendix B in Bankamp et al. 2022, paper II.2) and shape the primary data basis for Bankamp and Muntermann (2022, paper I.1), Bankamp and Muntermann (2021, paper II.1), and Bankamp et al. (2022, paper II.2). In addition to this unstructured analyst data, the Institutional Brokers' Estimate System (I/B/E/S) offers a simple and structured method of retrieving analysts' opinions. Among other information, the I/B/E/S database contains price targets, earnings per share (EPS) forecasts, and recommendations (buy/hold/sell). The data in I/B/E/S is already structured and processed, making it easier for researchers to use. Table 3 provides an excerpt from I/B/E/S showing the 12-month price targets for the Apple stock. It also indicates when the forecast was published (activation) and when the forecast was replaced by a new forecast from the broker (stop). I/B/E/S is an important database for academics conducting research on sell-side analysts (Payne and Thomas 2003). Structured data on analysts from the I/B/E/S database are used in Bankamp (2022, paper II.3) as the primary dataset and in Bankamp et al. (2022, paper II.2) as the secondary dataset. The simpler data handling compared to unstructured analyst reports also allows for a larger sample size in Bankamp (2022, paper II.3), which is primarily based on I/B/E/S data.

Instrument	Broker	Target	Currency	Period	Activation	Stop
AAPL.OQ	CLEVELAND RESEARCH	141	USD	12	2020-10-22	2020-10-30
AAPL.OQ	PIPER SANDLE	135	USD	12	2020-10-22	2021-01-27
AAPL.OQ	LOOP CAPITAL	115	USD	12	2020-10-25	2020-10-30
AAPL.OQ	ATLANTIC EQUITIES	150	USD	12	2020-10-26	2021-01-28
AAPL.OQ	RAYMOND JAMES	140	USD	12	2020-10-29	2021-01-25

Table 3. Sample Excerpt From the I/B/E/S Database

As a second category, media data is utilized in this thesis. Media data is used in Bankamp (2022, paper II.3) to operationalize the theoretical construct of media coverage. Data from RavenPack is used for this purpose. This database compiles over 22,000 different sources (RavenPack 2021), thus providing a holistic overview of media coverage with respect to a topic or an entity (e.g., company or person). RavenPack is also widely used in business research (see e.g., Drake et al. 2014; K.-Y. Ho et al. 2013; Shi et al. 2016).

The third group consists of company-related data, including constituent lists of indices (for sample selection), price data, and accounting data (primarily used as control variables in this thesis). This data is obtained via Refinitiv Workspace, Refinitiv Eikon, and Refinitiv Datastream and is incorporated into the individual studies. In addition, more specific company-related data is also considered. This data includes the 10-K reports obtained from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) Database for the analysis in Metzler et al. (2021, paper II.4). The second source for text documents in Metzler et al. (2021, paper II.4) comprises conference call transcripts obtained from Refinitiv Thomson ONE. Like analyst reports, these are available in PDF format but are more structured, allowing for easier processing. The data on the management board’s composition is the last part of company-related data utilized in Metzler et al. (2021, paper II.4). Due to the high proportion of missing data in individual databases on board information, numerous databases are compiled (Boardex, Amadeus, and Crunchbase) and supplemented by manual internet research to improve data quality.

B. Studies: Individual Research Contributions

I. Research Area: Methodological Foundation

Research Area I builds a methodical foundation for the remaining thesis. In Research Area II, numerous text mining methods, such as sentiment analysis or text classification, are applied. While these methods are now widely used, and best practices have emerged, the concept of document similarity is less developed. The study of Bankamp et al. (2022, paper II.2) is largely based on this concept, which has hardly been applied in the field of finance and accounting (see Loughran and McDonald 2016). This also means that few recommendations or best practices exist that researchers can rely on. This gap is closed by Research Area I. At the same time, the findings allow for a well-founded selection of methods for subsequent studies presented in Research Area II. The purpose of this research area is to answer Research Question I.1.

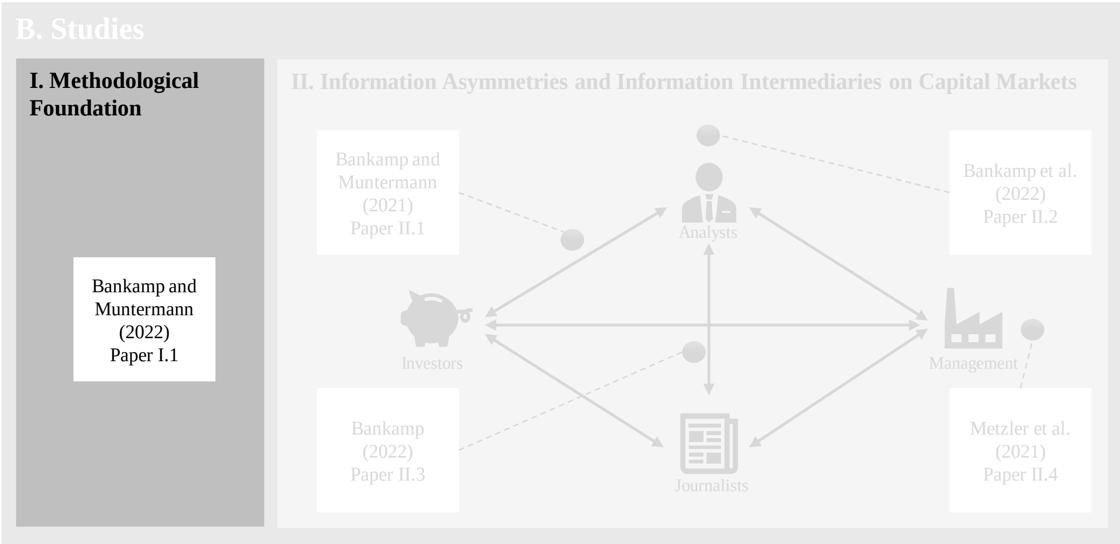

Figure 8. Overview of Research Area I

Research Question I.1: How should one choose document representations based on the dimension of similarity to capture?

I.1. Document Similarity

Understanding the Role of Document Representations in Similarity Measurement in Finance and Accounting

Authors: Steffen Bankamp and Jan Muntermann

Outlet: Proceedings of the 26th Pacific Asia Conference on Information Systems, Taipei/Sydney, Taiwan/Australia, *2022*.

Abstract: Document similarity is an important concept for many research questions. It can be applied to trace information exchanged on the capital market. For similarity calculations, the document must be transformed into a vector (document representation). Researchers can choose from a variety of document representations. We review the finance and accounting literature and find many different practices for estimating document similarity but little guidance on how to choose the right approach. To address this gap, we propose a framework of three similarity dimensions (object, author, and time). Based on this framework, we conduct an experiment on a corpus of analyst reports to quantify the accuracy of the estimated similarity. Our results help researchers and practitioners to choose an appropriate document representation for their analysis. Doc2vec achieves the overall highest accuracy, while Latent Dirichlet Allocation performs well on the object dimension. Bag-of-words models achieve surprisingly promising results despite their simplicity.

Keywords: Document Representation, Similarity, Finance, Accounting.

1 Introduction

Applying methods from natural language processing (NLP) in the finance and accounting domain is relatively new (Loughran and McDonald 2016). In contrast to standard statistical methods applied to structured data, hardly any conventions have been developed. At the same time, the analysis of numerous text documents in finance and accounting offers much potential to answer challenging and intriguing research questions (Loughran and McDonald 2016). The literature that has applied text mining methods has largely focused on sentiment analysis, while the application of document similarity remains underexplored in the finance and accounting domain (Loughran and McDonald 2016). However, document similarity can be a promising measure for important constructs, as it provides a direct measure of the informativeness of financial disclosures (e.g., Hanley and Hoberg 2010). It also allows for assessment of the time-varying association between companies (e.g., from annual reports) that extends beyond classical industry classification (Hoberg and Phillips 2016). Hence, overall document similarity is an important measure for constructs of interest and thus to test theories in the finance and accounting domain and beyond. Also for practitioners (e.g., asset managers or auditors), document similarity might help to filter for those documents containing new information and thus reduce the problem of information overload. However, before calculating these document similarities, the text must be transformed into a document representation, representing the document as a numeric vector of a specific length. Researchers and practitioners alike are faced with a large number of document representations from which to choose. Considering the early stage in the application of these methods in the finance and accounting domain, it is not surprising that hardly any conventions or best practices have been developed in the selection of document representations. Thus, little guidance is provided to authors faced with this decision. Only in Rawte et al. (2021), a comparison of different methods for text similarity estimations in the finance and accounting domain could be found. However, the authors only consider the temporal dimension of similarity and do not compare the accuracy of different representations for document similarity estimations. Knowing the accuracy of their methods is crucial for researchers and practitioners to make an informed decision. We address this issue and provide holistic guidance in this challenging selection process by answering the following research question:

RQ: *How should one choose document representations based on the dimension of similarity to capture?*

To answer this question, we first develop a framework encompassing three essential dimensions of similarity (*object*, *author*, and *time*). Based on previous research in the finance and accounting domain, we show that this framework is suitable for classifying the existing research. An experiment is conducted by analyzing over 200,000 financial documents (analyst reports) to answer the research question. We contribute to existing literature by providing methodological guidance to researchers and practitioners who want to calculate similarities between finance-related documents. This should support the selection of suitable methods. Our results suggest that irrespective of the utilized document representations, the object dimension is the most accurate and the temporal dimension the most difficult one to capture. Among the representations, doc2vec proves to be highly accurate across all dimensions. Surprisingly, simple methods such as term frequency (TF) and term frequency–inverse document frequency (TF-IDF) achieve promising results. In addition, the topic model Latent Dirichlet Allocation (LDA) provides high accuracy in detecting documents about similar companies (object dimension).

2 Theoretical Background

2.1 Document Similarity

The automated calculation of document similarity (sometimes also referred to as text similarity) is an important task within NLP (Shahmirzadi et al. 2019). The information generated from this calculation can be used in many applications, such as search engines (Pradhan et al. 2015) or question–answer matching (M. Tan et al. 2016). Document similarity might be applied as an operationalization for different constructs. Bär et al. (2011) argue that text similarity lacks a precise definition and that it is necessary to define what to measure with text similarity. The authors suggest three dimensions of text similarity: *structure*, *style*, and *content*. While we agree with Bär et al. (2011) that the question of similarity cannot be answered without a precise definition, the application of their dimensions requires manual coding to evaluate different methods of text similarity. To overcome this issue, we propose an approach where attributes for each dimension can be drawn from the metadata of many texts in the finance and accounting domain and beyond. We propose the dimensions *author*, *time,* and *object*, as illustrated in Figure 9. This allows for an evaluation of document representation for similarity measurement without manual labeling and thus an evaluation on a much larger dataset. In addition, the results are independent of human judgment, which would not be the case with a manually labeled dataset. It should be noted that the dimensions of Bär et al. (2011) and the three dimensions applied in this paper are related. The *author* dimension might capture elements of *style* and *structure*. This dimension can also be linked to the problem of author identification that is intensively discussed in the literature (e.g., Madigan et al. 2005). The *time* and *object* dimensions might relate to the *content* dimension of Bär et al. (2011). In our review of finance and accounting literature, we show that the dimensions we propose are actually present in this domain (see Table 4).

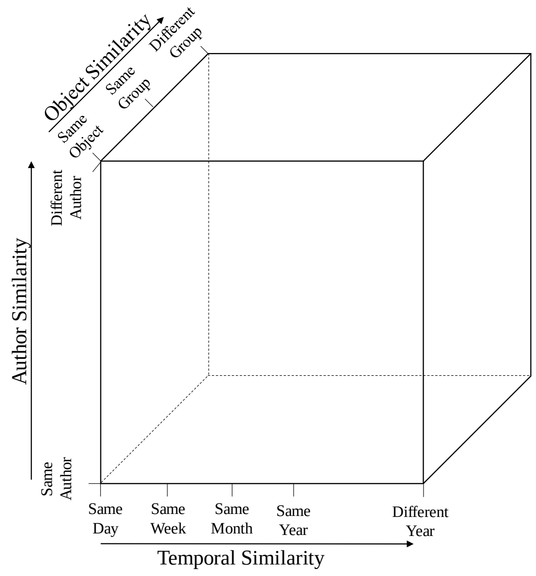

Figure 9. Similarity Cube

Our similarity framework is illustrated by a simple example. Consider the market for scientific textbooks. For each book, the three metadata characteristics discussed above should be available. If a retailer wants to suggest another book to a reader based on a book previously bought, the retailer could recommend books covering a similar object. In this case, the retailer must calculate the object similarity to all other books. However, the reader may not want to read any more books about this topic but likes the properties of the author (e.g., writing style) and would thus be interested in author similarity. Retailers that want to recommend the purchase of a book's new edition to a customer might be interested in temporal similarity. This example shows that not all similarities are the same, and substantial differences exist between the dimensions of similarity. At the same time, this also means that there is not necessarily a single operationalization to provide the best result for all these different constructs.

The technical process of calculating similarity can be divided into three steps, as illustrated in Figure 10. In the first step, the texts are pre-processed; this includes, for example, the transformation of the characters into lower case and stemming or lemmatization to reduce differently conjugated words or plurals to their basic form. The optimal extent of pre-processing can vary and depends on the dataset and the downstream analysis (Naseem et al. 2020). After pre-processing, the text is transformed into vectors, the so-called document representation. This transformation can be done using numerous different methods, which are discussed in detail later. Aside from their differences, the texts are always transformed into vectors of the same length. TF, for example, produces large and sparse vectors (several thousand elements), while document embeddings are dense and relatively small (usually 300–600 elements). The uniform shape of the vectors is necessary to compare the vectors of two documents. For this comparison, a similarity measure is applied that calculates the distance between the two vectors. Many different distance/similarity measures exist, some of which are shown in Figure 10. The result of this process is a numeric value that indicates how similar two texts are. The focus of this work is not on the whole process but on the choice of document representation. We apply the cosine similarity in our experiment, which is the predominantly used measure for text similarity (Singhal 2001). It should be noted that the choice of the similarity measure is less important, as A. Huang (2008) found similar results for a variety of similarity measures except for the Euclidian distance, which is less suitable for text mining.

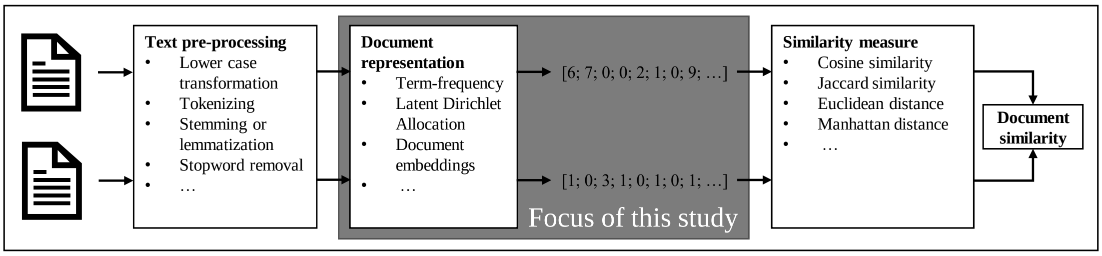

Figure 10. Pipeline for Similarity Calculation of Document Pairs

2.2 Document Representations

2.2.1 Bag-of-Words Models

Bag-of-words models are the simplest document representations. The document is represented only by the words it contains, and the order of the words is not considered. The length of the vector corresponds to the size of the corpus' vocabulary. As a result, the vectors can grow to a considerable size. Bag-of-words models also lead to sparse vectors (many elements containing zeros). Each feature represents an individual word. Features can be binary and thus only indicate whether the specific word occurs in the document or they can be a numeric feature (e.g., TF) and thus indicate the number of occurrences of the word within the document (Jurafsky and Martin 2009). An extension to this representation is TF-IDF. With this representation, the TF is normalized by the document frequency, which assigns more weight to words that occur only in a few documents and thus to more document-specific words (Sparck Jones 1972).

2.2.2 Topic Models

Topic models apply dimension reduction of the term-document matrix (TDM). Thus, as in bag-of-words models, the sequence of words is not considered (Blei et al. 2003). One approach to dimension reduction is latent semantic indexing (LSI) or latent semantic analysis (LSA) (often used synonymously). This approach was developed by Deerwester et al. (1990). The feature reduction is conducted by applying singular-value decomposition on the TDM. Words with semantic similarity are placed close to one another in the semantic space and are represented by the same topic (Deerwester et al. 1990). LDA is a probabilistic model that was developed by Blei et al. (2003).

Each document is considered as a distribution of topics and each topic in turn as a distribution of words. The advantage of LDA compared to LSA is higher generalizability, easier application to new documents, and less overfitting (Blei et al. 2003). The number of topics and thus the size of the vector constitute a central hyperparameter in both topic models discussed.

2.2.3 Word and Document Embeddings

Word embeddings represent a single word in a dense vector of defined size. The embeddings are derived from large, unlabeled text corpora. The word2vec embeddings developed by Mikolov et al. (2013a) are built either by predicting a current word from its surrounding words (continuous bag-of-words model) or by predicting surrounding words from the current word (continuous skip-gram model). This task is only performed to obtain the weights learned within the neural network for each word in the corpus. These weights represent the word embeddings. Global Vectors for Word Representation (GloVe) is a commonly used word embedding but uses a slightly different architecture. These embeddings are trained on the co-occurrence of words (Pennington et al. 2014). To convert word embeddings into a document representation, they must be aggregated to the document level. One possibility is to average all word embeddings of the words within the document to obtain a document representation (e.g., Mauritz et al. 2021). Kusner et al. (2015) suggest a word mover distance, which defines the sum of the shortest distance between the embeddings of all words from one document to another document. Alternatively, the user can directly apply document embeddings instead of word embeddings. Doc2vec (also known as paragraph vector), developed by Le and Mikolov (2014), is based on word2vec and uses a similar architecture. This method generalizes the word2vec architecture to the document level and directly provides document embeddings.

2.2.4 Transformer-Based Models

The invention of the transformer architecture (Vaswani et al. 2017) has enabled the development of many new language models. These include the bidirectional encoder representations from transformers (BERT) (Devlin et al. 2019) or the universal sentence encoder (USE) (Cer et al. 2018a). These models are much more complex than the previously discussed models and consider the position of words and their context within the document. In addition, these models are typically trained on several tasks, which leads to higher generalizability and state-of-the-art performance (Cer et al. 2018a; Devlin et al. 2019). However, the training of these models is much more complex than the training of the models discussed above.

2.3 Document Representation for Similarity Measurement in the Finance and Accounting Literature

Document representations can be applied for various downstream analyses beyond text similarity calculations. Yan et al. (2018), for example, use averaged word embeddings to build a classifier for loan project recommendations on social lending platforms. However, we only focus on literature that specifically applies document similarities. To determine how document representations are used in finance and accounting research for similarity calculations between documents, we analyzed the related literature. The results can be found in Table 4.

The use of document similarity as a variable in accounting and finance research is relatively new and not as established as sentiment analysis (Loughran and McDonald 2016). In their literature review, Loughran and McDonald (2016) identified three papers that use TF or TF-IDF as document representation for similarity calculations. They had already highlighted LSA as a promising document representation but could not identify any paper in the domain at that time. This is in line with our analysis of the literature, as the oldest paper we could identify that uses a document representation other than TF or TF-IDF for text similarity was published in 2018 (see Table 4). It took some time before these more advanced methods found their way into finance and accounting research. Apart from the application for text similarity, some document representations have been used earlier (e.g., Eickhoff and Muntermann 2016).

	Object	**Author**	**Temporal**
TF	(Hanley and Hoberg 2010) (Lang and Stice-Lawrence 2015) (Hoberg and Phillips 2016)	(Hanley and Hoberg 2010)	(Hanley and Hoberg 2010)
TF-IDF	(Brown and Knechel 2016)	(Mauritz et al. 2021)	(Brown and Tucker 2011)
LSA/LSI		(Beaupain and Girard 2020)	(Beaupain and Girard 2020)
LDA		(Palmer et al. 2018) (Mauritz et al. 2021)	
Word Embeddings	(R. Liu et al. 2020)	(Mauritz et al. 2021)	(R. Liu et al. 2020) (Adosoglou et al. 2021)
Document Embeddings			(Adosoglou et al. 2021)
Transformer	(J. Chen and Sarkar 2020)		(J. Chen and Sarkar 2020)

Table 4. Document Representation in Research of the Finance and Accounting Domain

The work of Hanley and Hoberg (2010) is a milestone in the field of financial and accounting document comparisons. The authors examine initial public offering (IPO) prospectuses to determine whether a higher proportion of informative content reduces

IPO underpricing. To distinguish informative content from standard phrases, they compare the document similarity between different IPO prospectuses. They demonstrate that IPO prospectuses from the same underwriter (author dimension), from companies in the same industry (object dimension), and that are published in a close temporal context (temporal dimension) are more similar than pairs of different underwriters, industries, or time frames. Thus, all three dimensions of similarity are addressed in this paper. The work of Brown and Tucker (2011) is also among the first in this field. They focus on the temporal dimension to show how the information value of the management discussion and analysis (MD&A) section of an annual report changes over time. Therefore, they compare the similarity between an MD&A section and the same company's MD&A section of the previous year. A high degree of similarity corresponds to low information value. The paper of Mauritz et al. (2021) is interesting, as the authors explicitly choose and justify different document representations for different constructs. They examine the auditor's role in the preparation process of financial reports and compare the financial reports of different companies that are audited by the same auditor. They found a higher degree of document similarity between annual reports audited by the same auditor compared to pairs of reports audited by different auditors. Even though the auditor is not the official author of the annual report, it is clear that Mauritz et al. (2021) focus on the author dimension. Interestingly, Mauritz et al. (2021) use three different document representations to measure different aspects of similarity (TF-IDF for wording similarity and LDA/word embeddings for content similarity). We could not find more complex document representations, such as document embeddings or transformer-based representations, used for similarity measurement in the classical accounting and finance literature. The papers using these advanced methods that we could identify tend to originate from computer science but address a problem from the accounting or finance domain. Adosoglou et al. (2021), for example, propose a system that uses document embeddings (doc2vec) to identify companies with little change in the semantic of their annual report compared to that from the previous year (temporal dimension) and show that abnormal returns can be achieved by investing in these companies. Overall, it can be concluded from the analysis that many different document representations are used in the finance and accounting literature. Our literature review also suggests that researchers measure different dimensions of similarity, sometimes even within the same study. The fact that all papers could be easily grouped into the three similarity dimensions proves the suitability of our proposed framework from Figure 9.

3 Data

To experimentally investigate the role of document representations for the similarity calculation in the finance and accounting context, we use a large sample of analyst reports from the Thomson ONE database. These are documents published by brokerage houses that analyze a company based on different approaches. Analysts usually provide price or earnings forecasts and make recommendations on whether investors should buy the company's stock (A.H. Huang et al. 2018). Analyst reports are suitable for the study because many different authors conduct analyses on the same object. In addition, the reports are published throughout the year rather than only once a year, as is the case with 10-K reports. This allows for the similarity and its dimensions (see Figure 9) to be analyzed as comprehensively as possible.

We choose the constituents of the S&P100 index as a sample. By using a sample of large companies, we ensure sufficient analyst coverage. We choose a relatively long time horizon of 13 years ranging from 01/01/2007 to 12/31/2019. With this time horizon, we can cover many analyst reports and are able to make more generalizable statements. To build our sample, we start with all analyst reports available from Thomson ONE in this period on companies that have been a constituent of the S&P100 within our observation period. We remove automatically generated analyst reports, extremely short reports of less than 300 words, and reports with more than 50 pages (as they are usually industry reports) from the sample. Reports that are written in a language other than English are also dropped from the sample. We then eliminate duplicates, as these would distort the experimental analysis of similarity. This leaves a total of 207,445 analyst reports on 137 companies published by 367 brokerage houses. The analyst reports are available in PDF format, and we perform standard pre-processing steps to make them usable for further analysis – these steps include the elimination of boilerplate, disclaimers, graphs, and tables. We transform the remaining text to lower case, remove stop words, punctuation, and numbers. In addition, we use lemmatization to reduce the words to their original form.

4 Experimental Design

The experimental design is built upon the similarity cube (see Figure 9). To evaluate how well the different document representations are suited to capture the three similarity dimensions, we keep two dimensions constant while varying the third. We do this by walking along the 12 edges of the similarity cube. The experimental design is illustrated in Figure 11.

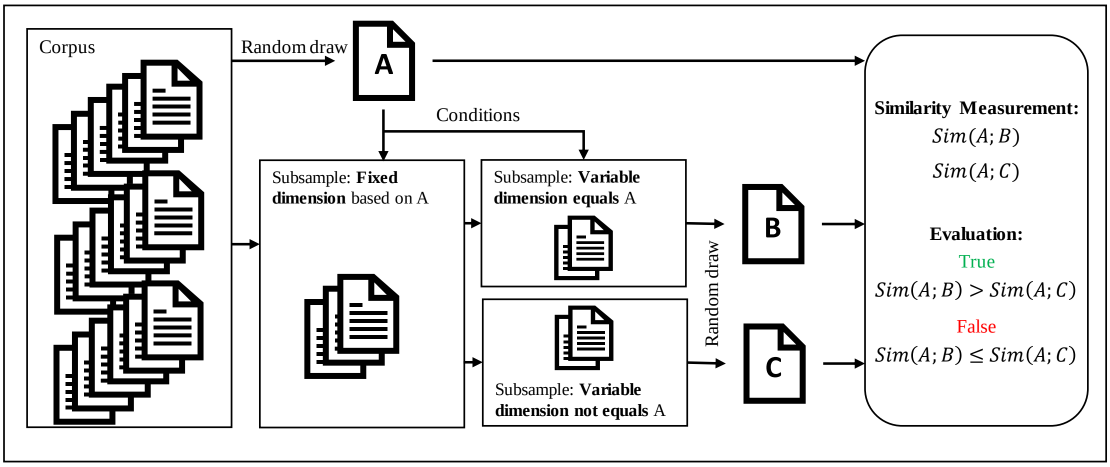

Figure 11. Experimental Design

As mentioned earlier, we avoid manual labeling by deriving the "true" similarity values from the metadata of the analyst reports. For each dimension, we categorize the similarity of report pairs into binary labels (similar vs. not similar). The categorization is presented in Table 5. The cutoffs for the temporal dimensions are derived from the analysts' observed publication structure. As analyst reports are strongly clustered on few days surrounding the earnings announcement (EA) (A.H. Huang et al. 2018), the bound of 14 days, on the one hand, ensures that these reports that are all linked to this date are actually recognized as temporally similar. On the other hand, the bound of 180 days ensures that the reports are at least two quarterly reporting periods apart and thus relate to a completely different point in time. We match groups consisting of three analyst reports. To assess how well the representations capture the object dimension under the condition of different authors (low author similarity) and the same publication time frame (high temporal similarity), we proceed as follows. We first draw a random analyst report from the entire corpus, which we call Report A. Subsequently, another report is drawn from a subset of reports about the same company as that of A (high object similarity) and are published by a different author at the same time relative to A. We call this Report B. Now we draw another report from a subset of reports about different companies than that of A (low object similarity) but with the same relationship to A in the other dimensions as Report B has to A (different object similarity and same temporal similarity). This report is called C. Now we calculate the similarity measures based on the different representations (see Table 6) between A and B and between A and C. If the similarity between A and B is higher than that between A and C, the representation has correctly captured the object similarity and the group is considered to be correctly classified. We form a total of 5,000 of these groups to obtain statically robust results. Reports that have been drawn are not reclined during the repeated drawings for the same combination of dimensions. The process is conducted for all 12 edges of the cube.

Dimension	Similar	Not similar
Object	Same company	Different company
Author	Same broker	Different broker
Temporal	Less than a 14-day difference	More than a 180-day difference

Table 5. Configuration of the Binary Experiment for Similarity Detection

Given the large number of different document representations and the even larger number of variations of these representations, a selection of representations must be considered for the experiment. The selection is based on the four overall classes of representations introduced in the theory section. For each of these classes, we choose at least one representation. The final selection within these classes is then based on the findings of our literature review (see Table 4). Thus, we predominantly select representations that are currently applied in the finance and accounting domain and that are popular among researchers and practitioners.

In total, we use eight different document representations, which are shown in Table 6, including their configuration. This selection covers a broad range of representations. For the transformer-based representation, the BERT model might be the most popular representation, and it is also the model applied in the paper of J. Chen and Sarkar (2020) (see literature review). However, we only consider representations that are actually capable of fully processing the data used in the experiment, otherwise it would not be possible to determine whether the measured effects are due to the representation itself or due to a partial recognition of data. As the BERT model has a limited input length (Sun et al. 2019), which many analyst reports exceed, we use the universal sentence encoder to represent the class of transformer-based models. This model does not have a constraint on the length of input data. We use a pre-trained model, since the training of such transformer-based models requires extensive computational resources and large text corpora. Furthermore, many researchers who want to apply these models might opt for the pre-trained models because of these demanding requirements. For the word2vec representation, we consider a model that is pre-trained on news articles and a model that we trained on our corpus of analyst reports. We also found both options in the literature we analyzed. Mauritz et al. (2021) use a pre-trained model, while R. Liu et al. (2020) train the model on their own corpus. All other representations are created based on our corpus of 207,445 analyst reports. The hyperparameters correspond to the typical values found in the literature or the default values of the software packages (see Table 6). Pre-processing is not applied to USE, as the required pre-processing is already built into the model's implementation (Cer et al. 2018a).

Representation	Configuration
TF	max doc frequency: 50% (Şaşmaz and Tek 2021) min doc frequency: 1% (González et al. 2015)
TF-IDF	max doc frequency: 50% (Şaşmaz and Tek 2021) min doc frequency: 1% (González et al. 2015)
LSA	max doc frequency: 50% (Şaşmaz and Tek 2021) min doc frequency: 1% (González et al. 2015) vector size: 100 (Deerwester et al. 1990)
LDA	hyperparameter: default from *mallet* vector size: 100 (Niraula et al. 2013)
Word2vec (news)	trained on news articles vector size: 300 (Jatnika et al. 2019) averaging of word vectors (Mauritz et al. 2021)
Word2vec (own)	hyperparameter: default from *gensim* trained on the corpus of the study vector size: 300 (Jatnika et al. 2019) averaging of word vectors (Mauritz et al. 2021)
Doc2vec	hyperparameter: default from *gensim* trained on the corpus of the study vector size: 300 (Trieu et al. 2017)
USE	TF2.0 Model (v4) no pre-processing (Cer et al. 2018a) vector size: 512 (Cer et al. 2018b)

Table 6. Configuration of Document Representation

5 Experimental Results

5.1 Three Dimensions of Similarity

As illustrated in Figure 12, significant differences exist in the accuracy of similarity recognition depending on the dimension combination and document representation. The x-axes of the bar plots show the configuration of fixed dimensions, and the y-axes indicate the accuracy of similarity recognition – the proportion of groups where Sim(A; B) > Sim(A; C). The horizontal dashed-dotted line indicates the expected value for a random estimation, and the error bars show the accuracy's 95% confidence interval.

The object dimension is easily detectable by most of the representations (upper left plot). This is hardly surprising, since the names of companies, products, or the management team provide features that should differentiate well between report pairs of the same vs. different companies. LDA, doc2vec, and TF-IDF perform particularly well. By contrast, word embeddings and the universal sentence encoder perform the worst. These types of models are designed to capture semantics and to generalize very well. However, these properties could also lead to difficulties in capturing the proper nouns described above, resulting in low performance of word embeddings and the universal sentence encoder on this dimension. This is especially the case for the pre-trained models, which may never have made contact with these proper nouns in pre-training.

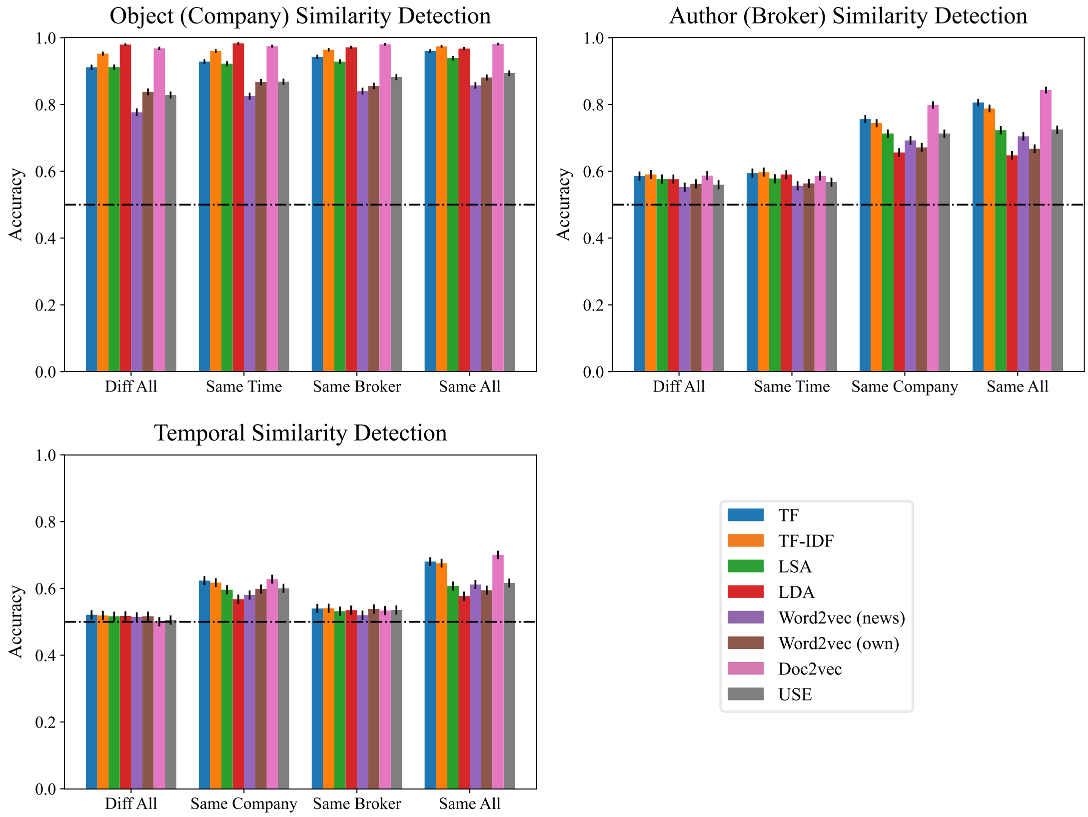

Figure 12. *Accuracy of Similarity Estimations Based on Various Document Representations*

The author dimension is more difficult to recognize (upper right plot of Figure 12) than the object dimension. Unlike the object dimension, where named entities provide features that contribute to an easier identification (e.g., the name of the company or its representatives that the report was written about), stylistic similarities play a more important role in the author dimension. These stylistic features are harder for the applied representations to capture. The identification of the author dimension is particularly difficult if different companies (objects) are involved. One explanation for this is that analysts working for the brokerage house are typically assigned to specific industries or companies. When examining reports from the same broker regarding the same company, it is more likely that the same analyst has actually written the article compared with report pairs that only share their broker. Doc2vec and the simple bag-of-words models (TF and TF-IDF) perform particularly well on the author dimension.

The temporal similarity dimension is the most difficult to capture. In the case of report pairs from different companies and brokers, the representations are hardly superior to a random estimation. This is not surprising, as only features such as the general economic environment or political decisions could indicate proximity. However, if the same company is considered, an estimation becomes significantly more accurate. Features that relate to the company's strategic actions are informative here. The sparse representations (TF and TF-IDF) and doc2vec perform well on this task.

Overall, the simplest document representations (TF and TF-IDF) perform relatively well. Doc2vec is one of the best-performing representations across all combinations. However, average word embeddings do not capture similarities particularly well in this experiment. It is also unclear whether a word2vec model trained on the data performs better than a word2vec model pre-trained on a general news corpus. Surprisingly, the modern USE method performs relatively poorly. However, this can be attributed to the fact that it was developed for short texts (especially sentences). Moreover, it was not trained on the corpus. The preceding analysis provides an initial overview of the similarity dimensions and representations. However, the dimensions have only been considered in a binary way (see Table 5). We extend this analysis and investigate the object and temporal dimension in more detail. Therefore, we examine the extent to which gradations within these dimensions can be recognized. For the broker dimension, we refrain from this analysis because no meaningful gradation for the similarity of the authors can be drawn from the available metadata, as a pair of analyst reports is published either by the same or by different brokers.

5.2 Detailed Analysis of Object Similarity

Most representations could identify relatively well whether a text pair concerned the same or a different company (see Figure 12). We attribute this to company names and other company-specific proper nouns. For practical applications, however, it is also important to recognize how similar the two companies are. To investigate this, we use The Refinitiv Business Classification (TRBC) to obtain the “true” values for company similarity. This classification consists of five levels: the highest level is the economic sector (e.g., Energy), and the lowest level is the activity (e.g., Wind Systems & Equipment) (Refinitiv 2022). Figure 13 illustrates how the average cosine similarity of pairs changes depending on whether the pairs share the company, the activity, or belong even to different economic sectors. The actual cosine similarity is shown in the left plot, and the *z*-transformed cosine similarity is depicted in the right plot. The *z*-transformation removes the different levels in mean and variance from the cosine similarity of the eight document representations. Each pair consists of reports from two different authors that were published within a time frame of 30 days.

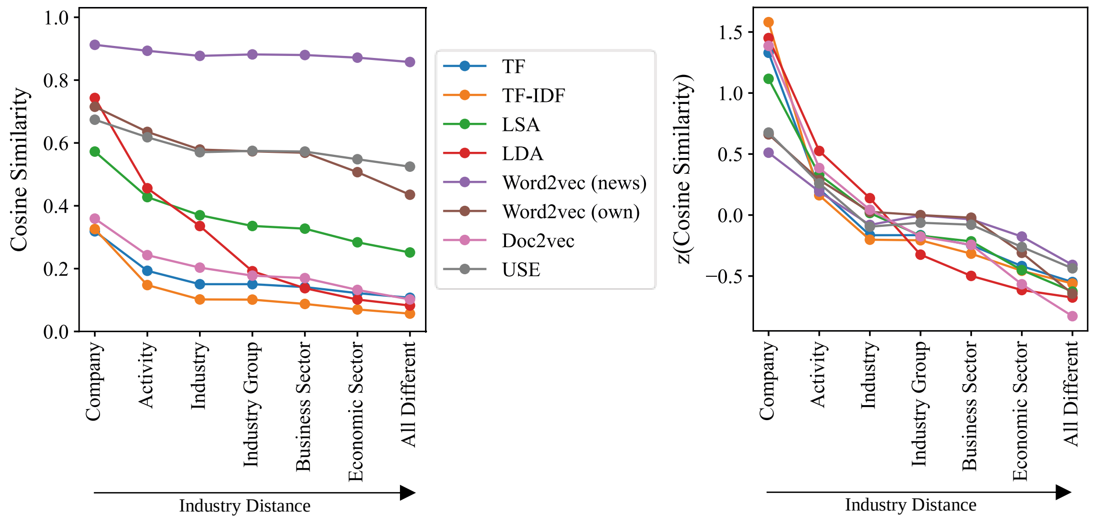

Figure 13. Document Similarity and Industry Distance

For all document representations, we observe a clear decrease in document similarity when the distance of the companies according to the TRBC business classification increases. The sharp increase from report pairs of the same activity to report pairs of the same company can, at least to some extent, be allocated to the proper nouns discussed above. This is also confirmed by the fact that the increase is much smaller in models such as USE or word2vec. LDA shows a strong differentiation between the industry, activity, and company level. This is also a reason why, in some studies, topic models are created for every single industry (A.H. Huang et al. 2018) or even for every company (Palmer et al. 2018). This prevents LDA from only representing industries or companies.

To quantify the representations' usefulness for the object dimension, 15,000 groups are formed following the experimental design (see Figure 11). Reports A and B now share their business activity but originate from different companies and the same time frame, and Reports A and C come from different business sectors. The calculation of accuracy is identical to those applied in Figure 12. Furthermore, Figure 14 indicates that LDA and doc2vec are still delivering the best results on the object dimension; however, the level is significantly reduced. Whereas LDA detected reports of the same vs. different companies (fixed dimensions: different author; same time frame) in 98% of cases (see Figure 12), report pairs on companies with the same activity are only detected correctly in 85.55% of cases (see Figure 14). Doc2vec achieves slightly but significantly better results (87.39%). Finally, TF-IDF is the third-best representation on this task, achieving an accuracy of 82.02%.

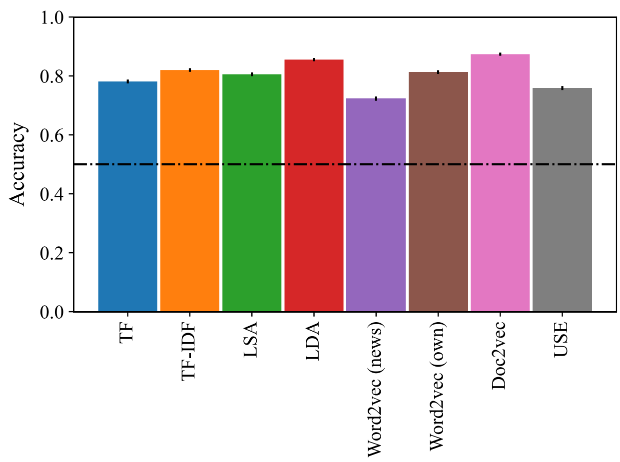

Figure 14. *Accuracy of Different Document Representations with Respect to Object Similarity*

5.3 Detailed Analysis of Temporal Similarity

The finance and accounting literature has especially focused on the temporal dimension of document similarity and applied a variety of representations to this task (see Table 4). To refine the binary analysis carried out earlier, report pairs are formed with different publication intervals ranging from 0 to 52 weeks. All pairs are built from reports on identical companies and different authors. The line plot (see Figure 15) shows how the temporal distance of pairs is related to document similarity. Unlike with object similarity investigated in Figure 13, we do not observe monotonously falling curves but a striking pattern of waves with a wavelength of approximately 13 weeks. This can be attributed to the quarterly earnings releases, which are important events for financial analysts and are discussed intensively in their reports (A.H. Huang et al. 2018). A report published shortly after the EA is more likely to be similar to a report published in 13 weeks than to a report published in two weeks and thus between two EAs. Such seasonal aspects are not specific to analyst reports but can be applied to many other domains. It is likely that seasonal fluctuations in the document similarity of news articles due to factors such as weather, holidays, or annual events are also observable. Since such a confounding factor can distort analyses, researchers should be aware of this problem and control for it. We do this by calculating the relative distance to the closest EA date for each report. We then form pairs where Reports A and B are published within five days, and Reports A and C are at least 70 days apart. However, both pairs must be published at the same time relative to the next earnings date. This procedure controls for the problem of seasonality. The report pairs are also about the same company and from different brokers. The results are depicted in Figure 16.

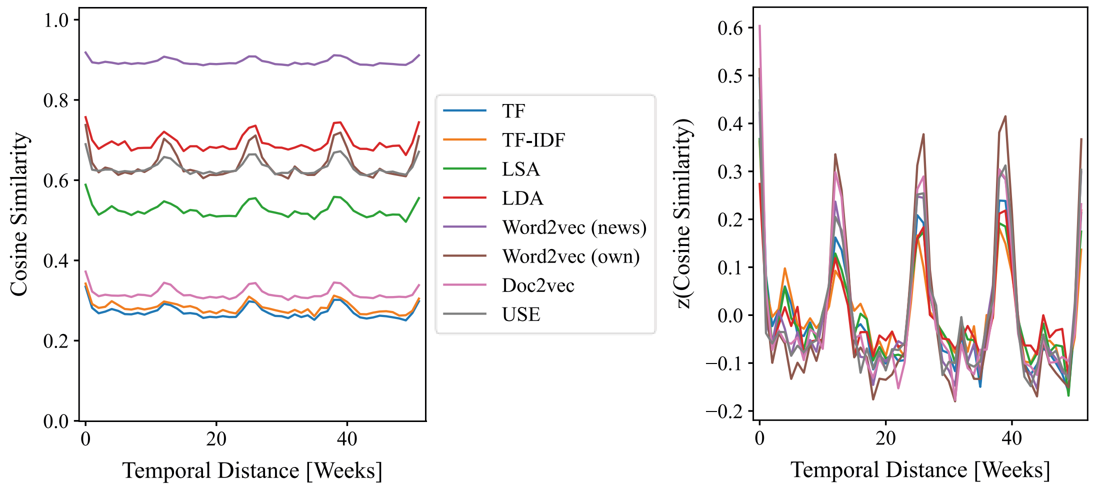

Figure 15. *Document Similarity and Temporal Distance*

It is revealed that doc2vec is again the most precise representation with an accuracy of 62.32%, followed by TF with 61.89%. The results are on the same level as those from the comparable analysis in Figure 12 without considering seasonality. LDA performs worst, with only 57.64% correctly identified pairs on their temporal proximity.

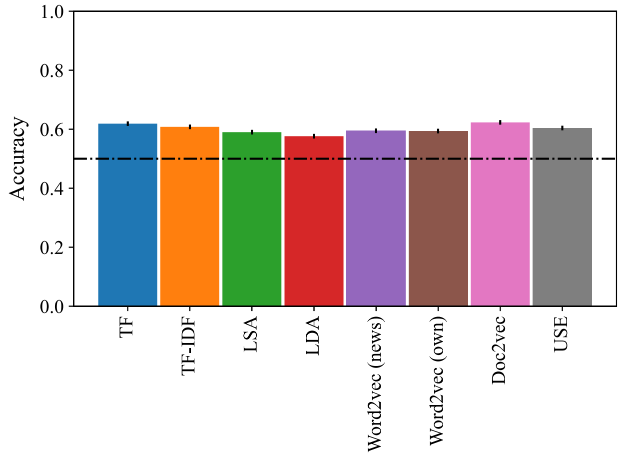

Figure 16. *Accuracy of Different Document Representations with Respect to the Temporal Similarity*

6 Discussion

Our paper provides numerous insights into the computation of document similarities in the finance and accounting domain. The similarity cube provides users with a useful framework to define the similarity dimension of interest. The analysis of the existing literature further suggests that these dimensions are well chosen and can be easily applied to the problems and research questions found in the literature. However, it seems plausible that this framework can also be applied to other domains, since the dimensions *author*, *object*, and *time* are generic. We also demonstrate that existing literature often applies certain representations without discussing how the representation fits the underlying research question. Only in Mauritz et al. (2021), we find that different representations were chosen for certain aspects of similarity. Thus, a systematic investigation of document representations for the estimation of document similarity seems appropriate.

Doc2vec has proven to be a prime general-purpose solution across all dimensions. Therefore, especially those who are not interested in measuring a specific dimension should consider using it. Moreover, LDA has proven to be particularly useful for recognizing the object dimension. To create text-based company clusters or peer groups, LDA could be a promising approach. Considering the other dimensions, however, LDA is not suitable. The dense and widely used document representations TF and TF-IDF also perform generally well on all three dimensions. This finding can help researchers who must balance accuracy on the one hand with simplicity and replicability on the other. The relatively poor performance of the universal sentence encoder in this study shows that the most complex models do not always lead to the most accurate results. This analysis can also help researchers and practitioners to detect and analyze seasonal effects in the studied document streams. Since these effects are likely to be present in many areas and are not limited to financial documents or analyst reports, they can distort analyses.

Our analysis comes with some limitations. First, as the study relies on a corpus of analyst reports, only one type of financial documents is used. This limits the generalizability of our analysis. In addition, the author dimension is based on the broker level. In fact, many analysts work for a single brokerage house, which is why relying on the analyst level would be more accurate. However, this is not used because analysts are industry or even company experts, which would result in a large overlap between the object and author dimensions. This problem is avoided by using the broker level. In addition, information about the analysts is not available in the dataset's metadata. Another limitation is the isolated consideration of the document representation. In real-world research projects, the choice of pre-processing, document representation, and similarity measure is not made independently. We make an exception in the case of USE and do not perform pre-processing for this representation (Cer et al. 2018a), but it would also be possible to improve the results of the other representations by adjusting pre-processing steps. For example, certain expressions could be concatenated by phrase detection (e.g., cash flow → cash_flow) (Mikolov et al. 2013b). Another possibility to improve performance, especially for LDA on the temporal and author dimensions, could be to create industry- or company-specific models (Palmer et al. 2018; A.H. Huang et al. 2018). For the USE representation, fine-tuning based on a classification task could be performed to potentially increase the performance. However, this would require researchers who want to apply the same approach to have a labeled dataset at hand. It would also mean that we compare a supervised learning method with unsupervised representations. Thus, we refrained from all these possible optimizations to avoid complicating the analysis by adding further combinations. Finally, it should be noted that neither the similarity dimensions nor the representations included in the analysis are exhaustive.

This paper offers multiple starting points for future research. On a conceptual level, the similarity cube can be supplemented by additional dimensions. An in-depth analysis of the interactions between the proposed dimensions would be an interesting task for future research. This could involve simultaneously changing several dimensions. At the same time, further representations should be evaluated using the framework and experimental design developed in this study. Since the calculation of the document similarity as shown in Figure 10 is complex and consists of several process steps, the neighboring process steps (text pre-processing and similarity measure) should be evaluated. The analysis of different combinations of the individual sub-steps could also be an interesting task for future research but will probably lead to high complexity. Furthermore, the document representation itself offers interesting opportunities for extension. These include enrichment with information from named-entity recognition (Friburger et al. 2002), which could be of particular interest in the finance and accounting domain. Table 6 shows that document representations require a comprehensive set of hyperparameters that can be tuned to achieve more accurate results than in our experiment. Future studies might guide researchers and practitioners to find an appropriate hyperparameter configuration depending on their problem. Future research should also address the issue of document length. In this study, we used very long documents, as they are common in the finance and accounting domain (e.g., 10-K or sustainability reports). However, this prevents the usage of models such as BERT, whose input length is limited (Sun et al. 2019). It is important to investigate how well the different models perform with short texts.

However, particular emphasis should be placed on future empirical research in the finance and accounting domain that applies document similarity to close research gaps. Whenever associations between objects, authors, or the temporal dimension are to be captured on a textual level, or when information flows are studied, the use of document similarity might be a methodological approach to consider.

7 Conclusion

Researchers and practitioners face the challenge of choosing from many available document representations and justifying their choice when calculating document similarity. In the finance and accounting domain, and probably beyond, researchers aim to capture many different constructs by using the similarity between documents. A review of the literature has revealed that there are no generally accepted best practices for choosing document representations. This applies to the overall level as well as to individual similarity dimensions. Our results suggest that, on the one hand, the use of doc2vec provides accurate results across all dimensions. The topic model LDA, on the other hand, accurately captures the object dimension but does not provide satisfactory results for the other dimensions. Furthermore, the simple and understandable bag-of-words models perform surprisingly well. In addition, we show that seasonality plays an important role when investigating the temporal similarity dimension and that it should be controlled for, where appropriate.

This paper is intended to help researchers from the finance and accounting domain when deciding which methodologies to use for answering interesting research questions that require a quantification of document similarity. The aim of this study is also to stimulate this kind of research, as the use of NLP in the finance and accounting domain is still largely focused on sentiment analysis and topic modeling (Kang et al. 2020; Loughran and McDonald 2016).

Document similarity provides an exciting playground for future methodological research. Since the measurement of document similarity is a highly complex task, there are many parameters whose effects on the results and accuracy should be further explored.

II. Research Area: Information Asymmetries and Information Intermediaries on Capital Markets

Research Area II builds the main part of this thesis. Bankamp and Muntermann (2021, paper II.1) deal with the design of a system to extract sustainability information from analyst reports and thus reduce sustainability-related information asymmetries between the management of a firm and its investors. Information intermediaries are important agents on capital markets and can contribute to the reduction of information asymmetries. Bankamp et al. (2022, paper II.2) and Bankamp (2022, paper II.3) examine how the MiFID II regulation from 2018 affects information intermediaries' output. Aside from information intermediaries, management might reduce information asymmetries via signaling. Metzler et al. (2021, paper II.4) examine how a CDO affects the volume of digital transformation-related signals that might reduce information asymmetries with respect to digitization projects.

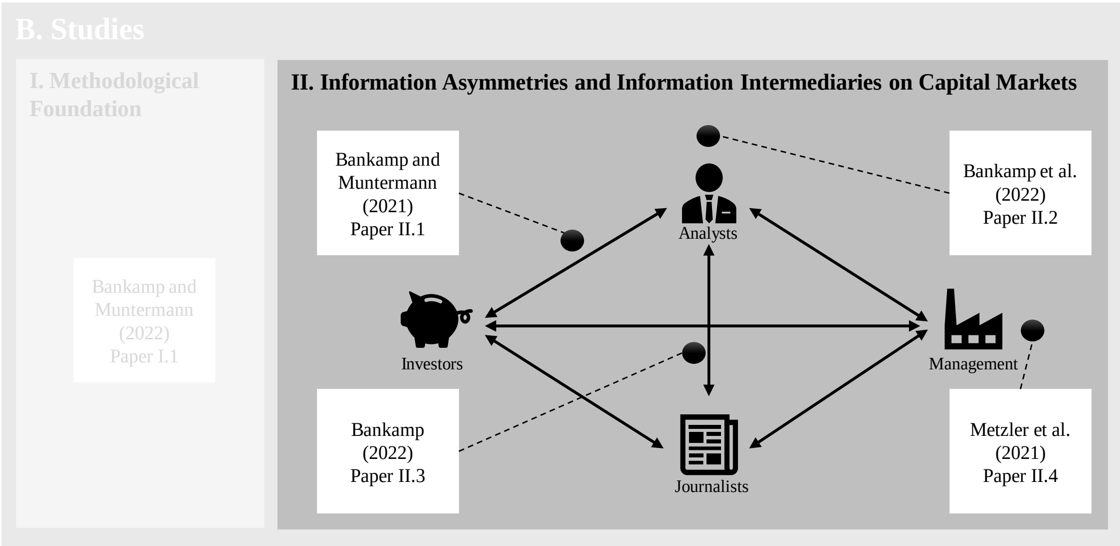

Figure 17. Overview of Research Area II

Answering the research questions below should contribute to a better understanding of the mechanisms that can reduce information asymmetries and current developments that might impact information asymmetries.

Research Question II.1: How should a system be designed for extracting sustainability-relevant information from analyst reports?

Research Question II.2: How do analysts modify their reports to respond to the changed market conditions induced by MiFID II?

Research Question II.3: How does MiFID II's research unbundling impact the provision of information by the media?

Research Question II.4a: How does CDO presence impact the volume of digital transformation-related signals in external communication tools?

Research Question II.4b: How does the volume of digital transformation-related signals differ across communication tools with different degrees of regulation?

II.1. Information Extraction on Sustainability

Are My Stocks Sustainable? Design Principles for Leveraging Information from Analyst Reports

Authors Steffen Bankamp and Jan Muntermann

Outlet Proceedings of the 25th Pacific Asia Conference on Information Systems, Dubai, United Arab Emirates, 2021.

Abstract: We address the problem of insufficient information about companies' sustainability, thereby helping investors to incorporate sustainability aspects into investment decisions. Building upon the design science research paradigm, we develop an artifact to extract information on companies' sustainability from text documents (analyst reports). We derive design principles that allow us to extract this information effectively and with a high degree of classification performance. The evaluation of the artifact shows that the proposed approach results in a precise extraction of sustainability-related information. Furthermore, this information is shown to be useful for supporting investors' decision-making.

Keywords: Sustainability, SRI, Design Science Research, Text Mining, Analyst Reports

1 Introduction

In 2015, the United Nations defined 17 Sustainable Development Goals (SDGs) whose implementation is intended to ensure global sustainable development at the economic, social, and ecological level (UN General Assembly 2015). Companies and their owners have a responsibility to contribute by their actions to achieve these goals. In recent decades, social and ecological indicators have become more important. Also, investors' awareness about sustainability issues has increased considerably (Flammer 2013). Consequently, there is strong growth in financial products that take sustainability issues into account, e.g., by excluding companies that do not fulfill their sustainability criteria (Global Sustainable Investment Alliance 2018). It is also apparent that fund managers rely increasingly on information about sustainability (van Duuren et al. 2016; Amel-Zadeh and Serafeim 2018). This has given rise to a new industry that conducts sustainability assessments and ratings and makes them available to investors. Simultaneously, more than 40% of institutional investors do not fully consider sustainability issues because it is too expensive to obtain and collect this information (Amel-Zadeh and Serafeim 2018). There are also data quality concerns with the Environmental, Social, and Corporate Governance (ESG) ratings produced by professional agencies (Kotsantonis and Serafeim 2019). These scores usually do not provide contextual information about incidents within the company. To make the economy more sustainable, investors can make an important contribution by directing capital to those companies that contribute to the achievement of the SDGs and, in return, withdrawing from unsustainable companies (Pástor et al. 2021). In addition to investors' willingness to invest in sustainable companies, the information base on which capital flows are allocated is crucial.

Against this background, our goal is to develop an artifact to extract sustainability-relevant information from analyst reports. These reports are prepared by financial analysts and distributed to investors. Financial analysts are important information intermediaries in the financial market and discuss a broad range of topics in their reports (A.H. Huang et al. 2018). Nilsson et al. (2008) show that analysts also discuss sustainability-related topics in their reports. However, this information is dispersed throughout the analyst reports, making a manual extraction time-consuming for investors. An automated extraction is necessary to make this information easily accessible to investors in an aggregated form. Due to the importance of analyst reports for investors, it can be assumed that these documents are already available to many investors, especially to institutional investors. The extracted information on sustainability should improve the decision-making of investors concerning the assessment of companies' sustainability. Hartzmark and Sussman (2019) show a shift in capital allocation from less sustainable assets to more sustainable assets when sustainability-related information is made easily accessible to investors. The provision of information is thus an elementary building block for achieving a more sustainable economy. To contribute to this goal, we follow the design science research (DSR) paradigm (Hevner et al. 2004) and apply the process model from Kuechler and Vaishnavi (2008) to develop the artifact. The artifact is evaluated with regard to the performance of classification and the informational value of the extracted information.

2 Research Background on Sustainability in Investing

The idea of taking ethical considerations into account when making investment decisions can be traced back hundreds of years. However, the strong growth in this field has just been observed for several decades and has increased with the general awareness of sustainability issues and with past environmental disasters (Schueth 2003). This type of investment is often called socially responsible investing (SRI) in the financial literature (e.g., Nofsinger and Varma 2014; Kempf and Osthoff 2007). Contrary to what the term suggests, not only social aspects are considered, but also other ethical aspects like environmental issues. According to Schueth (2003), there are three strategies an investor can choose from to implement SRI. First, screening is a strategy by which the investor reduces the investment choice set based on her ethical values. This happens most commonly through negative screening (Amel-Zadeh and Serafeim 2018), where specific companies or industries are excluded that do not fulfill the investor's minimum standards (van Duuren et al. 2016). With positive screening, the investor focuses specifically on sustainable companies or industries (van Duuren et al. 2016). The second strategy proposed by Schueth (2003) is shareholder advocacy. Investors are applying this strategy by influencing the management's decision-making (e.g., through their voting rights at the annual general meeting) to make the company more sustainable. The third strategy from Schueth (2003) is community investing, in which investors provide capital to weaker communities, thereby enabling the financing of low-income housing and small businesses. One major strand of literature within the field of SRI has analyzed the relationship between companies' financial and sustainability performance. Friede et al. (2015) were able to identify over 2,000 studies about this question. Overall, they found a positive correlation between financial and sustainability performance.

Following Hartzmark and Sussman (2019), investors prefer sustainable assets and react to new information regarding assets' sustainability by redirecting their capital. This is in line with the findings of Amel-Zadeh and Serafeim (2018), who found that even 82% of fund managers of conventional (non-sustainable) funds state to consider sustainability aspects in their investment decision. Pástor et al. (2021) show by means of an equilibrium model that these financial investment decisions have a positive impact, as real investments are shifted from non-sustainable to sustainable firms, which makes the economy more sustainable. At the end of 2019, sustainable assets accounted for 15.1% of the total assets held by mutual funds in Europe and are expected to increase to 41-57% by 2025 (PwC 2020). Thus, SRI represents a significant share of the capital market and has large growth prospects.

3 Methodology and Problem Description

3.1 Design Science Research

In order to mitigate the challenge of insufficient data on companies' sustainability (PwC 2020), we build upon the design science research paradigm (Hevner et al. 2004). In DSR, a solution for a problem is developed based on the current state of knowledge (e.g., theories, frameworks, methods). The artifact is developed during the DSR process and "*extend[s] the boundaries of human problem solving and organizational capabilities*" (Hevner et al. 2004, p. 76). DSR can contribute to existing knowledge by providing constructs, models, methods, instantiations, and design theories (Gregor and Hevner 2013).

In this study, we develop an instantiation that can be classified as improvement research according to Gregor and Hevner (2013). The artifact extends the problem class (sustainability data provision) by accessing a previously unexploited source of information on companies' sustainability. According to the belief-action-outcome framework of Melville (2010), our artifact should lead to action formation because it is intended to enable investors to consider sustainability issues during their investment-related decision making.

3.2 Research Process

We utilize the DSR process model proposed by Kuechler and Vaishnavi (2008). Starting with the *awareness of the problem*, we derive the problem of insufficient sustainability data from the existing literature. Building upon this, we identify three design requirements (DR) related to a potential problem solution. In the second process step (*suggestion*), we propose three design principles (DP). We derive concrete design features (DF) from our design principles and implement them by building the IT artifact in the *development* step. In the fourth step, the *evaluation* of the artifact is conducted in two stages. First, we evaluate the classification performance of the artifact based on different quantitative performance metrics. Finally, we qualitatively evaluate the usefulness of the extracted information for investors. In the final process step (*conclusion*), the acquired design knowledge is summarized and future research opportunities are presented.

3.3 Problem Description

Achieving the SDGs is a major task for society. As explained before, investors take sustainability aspects into account. This sustainability-guided allocation function of capital also leads to real investments (e.g., production facilities) being shifted from less sustainable to more sustainable companies (Pástor et al. 2021). These real investments can contribute to the achievement of the SDGs. However, successful SRI relies on available data on companies' sustainability. Only reliable and comprehensive data ensures an effective allocation of capital flows towards more sustainable companies and, in turn, into more sustainable real investments.

However, data about the sustainability of companies poses a significant problem. PwC (2020) found that for 73% of the surveyed asset managers, the lack of data is the largest barrier to implement sustainable products. Also, the high cost of aggregating sustainability information is a major obstacle for taking environmental data into account (Amel-Zadeh and Serafeim 2018). There are two major sources for ESG information. First, self-disclosures made by the companies in the form of Corporate Social Responsibility (CSR) reports, and second, ratings made available by specialized ESG rating agencies. Companies' self-reported data are less objective as negative aspects are inadequately reported (Chauvey et al. 2015). This data is found to be skewed and inaccurate, which questions the reliability and presume greenwashing (PwC 2020). Institutional investors criticize the nonspecificity of this information for using it in a targeted manner (Amel-Zadeh and Serafeim 2018). Designated ESG ratings suffer from data quality issues (Kotsantonis and Serafeim 2019). Also, sustainability principles as "Life-Cycle-Thinking" have not been integrated into ESG Rating agencies' assessments (Escrig-Olmedo et al. 2019). Simultaneously, there is a large dispersion between the assessment results of different rating agencies on the same company (Berg et al. 2020; Dimson et al. 2020; Kotsantonis and Serafeim 2019). This can be in particular attributed to different measurement methods. Berg et al. (2020) call for more transparent ESG ratings regarding their measurement methods. According to PwC (2020, p. 36), "*Traditional ESG data and ESG scoring will no longer suffice.*" Dimson et al. (2020) argue that ESG ratings should not be applied blindly but supplemented by the asset manager's own review.

4 Artifact Design

4.1 Design Requirements

In order to support the decision-making of investors in the selection of sustainable companies, we propose analyst reports as an additional source of information that can supplement the existing information sources (CSR reports and ESG ratings). Analyst reports provide important information to investors. They deliver financial analyses about public companies and contain investment recommendations. Financial analysts take the important role of information intermediaries on the capital market (A.H. Huang et al. 2018). The reports provided by analysts focus on financial and business aspects of the company. This can be seen from the fact that most of the analysts' reports are published in a narrow timeframe surrounding the quarterly conference calls and earnings announcements (A.H. Huang et al. 2018). Nevertheless, Nilsson et al. (2008) show that sustainability aspects are also discussed in analyst reports. It seems appropriate to leverage this information source as well. Compared to the two information sources described above, analyst reports have three advantages. First, unlike the CSR reports, the analyses are not written by the company itself but by a third party, enhancing objectivity. Second, in contrast to ratings from specialized ESG agencies, the analyst reports do not have to be acquired additionally but should already be available to most institutional investors as it is a common source for financial decision-making. Third, the analyst reports discuss sustainability aspects in textual form, which, in comparison to the frequently used quantitative rankings, allows investors to make informed decisions, considering their individual ethical principles and guidelines, and can thus better argue their decision to end-investors or other stakeholders.

As analyst reports cover primarily financial topics, the sustainability topics have to be extracted. Nilsson et al. (2008) found that for companies operating in the oil/gas and chemical industry 35% of the analyst reports contain environmental information. In other industries (e.g., semiconductor or telecommunication), the proportion is lower (Cerin 2010). There is also relatively little environmentally relevant information within a single report (Nilsson et al. 2008). Since sustainability-relevant topics seem to be only sporadically present in analyst reports, and this information has to be gathered across a large universe of companies, especially if negative screening (van Duuren et al. 2016) is conducted, *automated and precise extraction of sustainability-relevant information* is necessary. This builds DR1.

The sporadic content on sustainability (Nilsson et al. 2008) is embedded in a much larger part of financial information (A.H. Huang et al. 2018). As a result, there is a significant class imbalance in analyst reports between sustainability-related content and other content. Therefore, the related problem solution must have the *ability to handle extremely imbalanced datasets,* which represents DR2.

Based on the literature, we have identified that in today's ESG ratings, divergence among agencies is a major problem (Berg et al. 2020). At the same time, 43.2% of fund managers say that a lack of standards prevents them from considering ESG-related information in their investment decisions effectively (Amel-Zadeh and Serafeim 2018). Therefore, we consider DR3 elementary, as the extraction of information has to be *based on a common understanding of sustainability*.

4.2 Design Principles

Based on the three design requirements discussed above, we derive design principles and link them with related requirements (DRn→DPn). In order to ensure the automated extraction of sustainability-relevant information required by DR1, various methods are conceivable. First, it can be done based on dictionaries (wordlists). A text section is classified as sustainability-relevant if it contains one or more words from the dictionary. For example, the studies of Nilsson et al. (2008) and Cerin (2010) used predefined keywords to search for environmentally relevant content in analyst reports. Dictionary-based methods are also frequently used for sentiment analysis in the finance and accounting literature, where sentiment is extracted from texts (Loughran and McDonald 2016). The advantages of such a dictionary are that it can be applied straightforward once it has been created, it can be easily transferred to other documents, and it is replicable (Albaugh et al. 2014). A disadvantage of dictionaries is that it only recognizes the exact terms it contains. Dictionaries are not able to recognize synonyms unless they are already included in the dictionary. It is therefore particularly promising if there is a rigid terminology for the corresponding topic (Albaugh et al. 2014). Second, machine learning methods such as a support vector machine or a neural network can be applied for a classifier. The training of the classifier is conducted on a labeled dataset of text sections containing sustainability-relevant information and text sections not containing this information. A trained classifier can then be applied for inference, classifying unlabeled text sections automatically. A.H. Huang et al. (2014a) showed for the analyst domain that a sentiment classifier based on machine learning (naïve Bayes) is more accurate than a dictionary-based classification. However, a sufficiently large training dataset is required for training the classifier. Due to the expected class imbalance (DR2), it is questionable whether it is possible to manually label enough text sections containing sustainability-relevant information within a reasonable amount of time. In addition, DR3 requires that the sequence extraction is built upon a common understanding of sustainability. This can be achieved more adequately by using a domain-specific dictionary. In order to provide high extraction performance (DR1) and to address DR2 and DR3, we propose a combination of these two approaches. Related to all design requirements (DRn→DP1), we derive DP1, according to which a *hybrid approach that combines a dictionary with machine learning* should be followed. Eickhoff (2015) shows that dictionary-based and machine learning-based classifiers can be successfully combined in a hybrid approach.

To achieve a precise extraction (DR1), not only the sole word count as in bag-of-words models should be considered, but also the textual context. The finance and sustainability domain share many words that have different meanings. For example, the word "disposal" is commonly used in finance contexts to refer to the sale of parts of a company. In the context of sustainability, it instead refers to the discharge of waste. To address this problem (DR1→DP2), we are *considering the context in textual data* (DP2).

To further address DR3, we derive another design principle (DR3→DP3). To increase the generalizability, we rely on existing knowledge from SRI and from frameworks for sustainability reporting that we consider as kernel theory informing our artifact design. This ensures that the extraction of text sections is grounded on a broadly accepted understanding of sustainability and that the results are thus accepted by a majority of users. DP3 is the *theory-based design of the artifact that utilizes existing frameworks and knowledge from sustainability reporting*.

4.3 Design Features

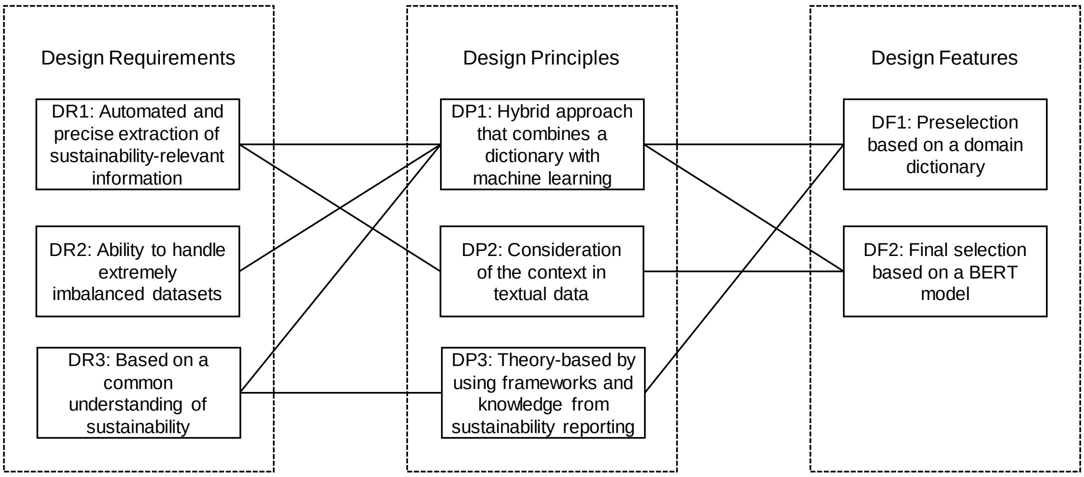

Figure 18. Design Requirements, Principles, and Features

Based on the three design principles, we derive and link artifact-related design features (DPn→DFn). The mapping between design requirements, principles, and features is illustrated in Figure 18. Design features describe the specific technical design of the artifact. This distinguishes design features from design principles, which describe the artifact properties at a higher level of abstraction (Meth et al. 2015). Incorporating the design features into the artifact should enable the artifact to fulfill the design requirements. As the first design feature (DP1/3→DF1), we define the *preselection of text* *passages using a domain dictionary*. We select the sentence level as the granularity of the text passages. A sentence is included in the preselection if this sentence contains at least one word or n-gram of the dictionary. The selection of the dictionary is an important parameterization of the artifact. A review of the knowledge base shows that two sustainability dictionaries have been developed. Deng et al. (2017) present a dictionary for environmental sustainability within the IT industry. However, since our approach is to develop an artifact that is not restricted to a specific industry, this dictionary does not fit. Pencle and Mălăescu (2016) have developed a dictionary on four CSR dimensions (employee, environment, human rights, and social community). This dictionary is based on a deductive (derived from literature) and an inductive (derived from IPO prospectuses) approach. The dictionary is general in terms of the industry but is also tied to the finance literature. Therefore, it seems to be adequate in the context of our problem class. This design feature (DF1) is the first module of the hybrid approach (DP1) and draws on existing sustainability-related knowledge (DP3).

To increase the precision of the sequence extraction, a *final selection of sequences is conducted by applying a Bidirectional Encoder Representations from Transformers (BERT) language representation model* (DP1/2→DF2). This approach replaces the manual evaluation done by Nilsson et al. (2008) following the keyword search. For this purpose, we build a binary classifier based on the BERT model (Devlin et al. 2019). This model consists of a deep neural network that is unsupervised pre-trained on the English Wikipedia and a large dataset of books (Devlin et al. 2019). The pre-training is based on a cloze task and a next sentence prediction. For the cloze task, the model is trained to predict masked tokens within a sentence. For this prediction, the entire sentence (except for the masked token) is available to the model, which is why it is called a bidirectional model. This is an important difference from so-called unidirectional architectures, where a word is predicted only from the preceding or following tokens. This allows the model to learn the entire context of a word. In addition, the next-sentence prediction is used in pre-training, where the model is trained to predict whether a chain of two sentences consists of consecutive sentences or not. The pre-training reduces the computational effort for task-specific training (fine-tuning) substantially (Devlin et al. 2019). In fine-tuning, a classification layer is added to the model. With a binary classification problem (sustainable/non-sustainable), this layer will have two outputs. In fine-tuning, the model is trained for the specific task based on the labeled dataset. However, with the BERT-architecture, not only the weights of the classification layer are adjusted, but also those of the entire model.

The BERT-model is suitable for many text mining tasks, including sentence classification making it suitable for the problem at hand. In contrast to bag-of-words models, it also considers the sequence within a sentence and contextual information (DP2). Furthermore, based on BERT, significantly better results on text mining tasks could be achieved than by prior methods (Devlin et al. 2019). We consider it useful to use BERT for the final selection of the preselected sentences and thus to be the second module of the hybrid approach (DP1).

5 Artifact Evaluation

5.1 Dataset

To evaluate the artifact, we use a comprehensive dataset of analyst reports. As a company sample, we select all companies of the Dow Jones Industrial Average (major US index) and the EuroStoxx 50 (major European index) that have been a constituent at any time during our investigation period ranging from 01-01-2015 to 12-31-2019. This results in a sample of 90 companies. This sample includes companies across a wide range of industries, including chemical and gas companies, for which a relatively large amount of information on environmental issues has been found by prior research in analyst reports, as well as companies in the telecommunications industry, for which significantly less information has been found (Cerin 2010). Both the selection of two large capital markets and the large variety of industries increase our study's generalizability. We collect all analyst reports available from Refinitiv Thomson ONE about the 90 companies during the investigation period, resulting in a sample of 95,665 reports. To clean the dataset, we remove duplicates, automatically generated analyst reports, analyst reports with more than 50 pages (typically industry analysis) and short updates with less than 300 words. This leaves 61,592 analyst reports as final sample. Table 7 shows the ten companies with the most reports of the two stock indices as well as the ten brokers who published the most reports. From the analyst reports (PDF files), we extract the text and remove charts, tables, diagrams, and boilerplate such as disclaimers. Since information extraction is done at the sentence level, the text is split into sentences. This results in 3,410,598 sentences that are used for further analysis. To develop and evaluate the instantiated artifact, we limit the scope to the extraction of information regarding the environmental dimension of sustainability. According to Hartzmark and Sussman (2019), the environmental dimension is most strongly associated with sustainability (79%). Despite the multifaceted nature of sustainability, we therefore consider it appropriate to first focus on the environmental dimension. For this reason, we use only the environmental word list of Pencle and Mălăescu (2016). This consists of 451 entries, including 323 uni-grams, 114 bi-grams, 10 tri-grams, and four entries consisting of more than three words.

Top 10 Europe		Top 10 US		Top 10 Broker	
Company	N Reports	Company	N Reports	Broker	N Reports
SAP	1,156	Apple	2,085	JP Morgan	5,323
Volkswagen	1,041	Intel	1,380	Morgan Stanley	4,323
AB InBev	914	Boeing	1,188	Morningstar	3,960
ASML	820	Walmart	1,177	UBS	3,755
Nokia	805	Microsoft	1,177	Deutsche Bank	3,682
Bayer	801	Caterpillar	1,171	RBC	3,293
Sanofi	767	Cisco Systems	1,156	Barclays	3,018
Airbus	760	Johns. & Johns.	1,105	Société Générale	2,819
Daimler	753	General Electric	1,060	Credit Suisse	2,689
Telefonica	741	Merck & Co	954	Jefferies	2,480

Table 7. Top Companies and Brokers within the Sample

5.2 Evaluation Results

In the first step, we evaluate whether the hybrid approach is even necessary. Alternatively, a training dataset could be labeled directly and the classifier trained based on this training dataset. For this purpose, 1,000 randomly selected sentences of the whole corpus were manually labeled. The labeling is based on the Global Reporting Initiative (GRI) framework (GRI 2020), which is the globally dominant standard in sustainability reporting (KPMG 2020). Many companies apply the GRI framework when preparing their sustainability report. The GRI Standards can be divided into series. The GRI 100 series define basic universal standards. The GRI 200, GRI 300, and GRI 400 series contain topic-specific standards. The GRI 300 series deals with environmental aspects, while GRI 200 focuses on economic and GRI 400 on social issues. Each series contains a set of standards. GRI 302-3, for example, defines the reporting of the organizations' energy intensity (GRI 2020). We classify a sentence as environmentally relevant if it contains an aspect of the GRI 300 series. Labeling this way ensures that the allocation is based on a common understanding of sustainability. Only three of the 1,000 randomly labeled sentences contained sustainability information. If this proportion of 0.3% ($CI_{0.95}$ = [0.11%; 0.87%]) corresponds to the true proportion in the overall sample, over 33,000 sentences would have to be labeled to obtain a sample containing 100 environmentally relevant sentences. Even this training dataset with 100 sentences of the minority class would still be relatively small. To solve this problem and generate a training and validation dataset more efficiently, we consider the hybrid approach to be appropriate. The hybrid approach is illustrated in Figure 19, which shows the process of prototype development (i.e., artifact instantiation) and evaluation.

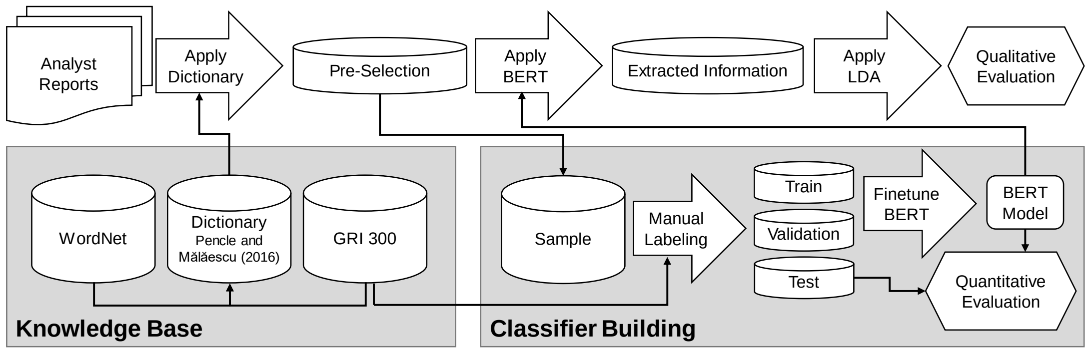

Figure 19. Process Map for Prototype Development and Evaluation

Pencle and Mălăescu (2016) developed their dictionary to evaluate the level of CSR language in financial documents. It is therefore not explicitly designed to use for the task of information extraction. For this reason, we adapt the dictionary to increase both sensitivity and precision in recognizing sentences with environmental content. To increase sensitivity, we first add 70 expressions (e.g., “fuel consumption”) that are central within the GRI 300 standards but are not part of the Pencle and Mălăescu (2016) dictionary. By adding relevant words to the dictionary, the likelihood of sustainability-related sentences being recognized by the dictionary is increased, which improves sensitivity. We further extend the dictionary by adding synonyms and lemmas of all elements from the dictionary by using the lexical database WordNet (Fellbaum 1998). This results in a dictionary containing 1,941 entries. To increase precision, we then exclude words or n-grams that cannot be directly assigned to the standards of the GRI 300 series. The excluded words are very generic (e.g., “accept”, “design” or “grow”) and are therefore not suitable for an information extraction system with a high degree of precision. By removing these generic words, the probability of false-positive results is reduced. This in turn improves precision. The remaining dictionary consists of 402 words or n-grams, where 341 (84.83%) entries can be mapped to a single standard within the GRI 300 series (e.g., “energy efficiency” → GRI 302). The remaining 61 entries have a meaning that covers several of the sub-standards.

We apply the dictionary to the whole dataset. 65,848 of the total 3,410,598 sentences (1.93%) are preselected by the dictionary. These preselected sentences contain at least one word from the dictionary. We then manually labeled 4,000 randomly selected sentences of the preselection. 752 (18.8%) sentences were identified as environmentally relevant. This shows the suitability of the preselection and that, despite less labeling, a relatively large training dataset in terms of the positive class could be obtained. The preselection rate (1.93%) and the positive rate in the preselection (18.8%) result in an expected value for the total extraction of 0.364%. Since this value is above the positive rate of 0.3% found in the initially labeled sample of 1,000 sentences from the population, the preselection does not lead to a substantial decrease of sensitivity.

The 4,000 labeled sentences are randomly split into a training (N=2,400), validation (N=600), and test dataset (N=1,000). The training dataset is used to fine-tune the BERT model. As this is done in epochs, the validation dataset is used to stop the learning process when the highest predictive performance based on the validation dataset is achieved. This procedure prevents the model from overfitting.

Based on the test dataset, the BERT model achieves a precision of 0.7573 (75.73% of the sentence classified as environmentally relevant are actually environmentally relevant) and a recall of 0.8342 (83.42% of sentences within the preselection that are environmentally relevant are classified accordingly). Different precision-recall combinations of the classifier (solid black line) and receiving operator characteristics (ROC) can be derived from Figure 20.

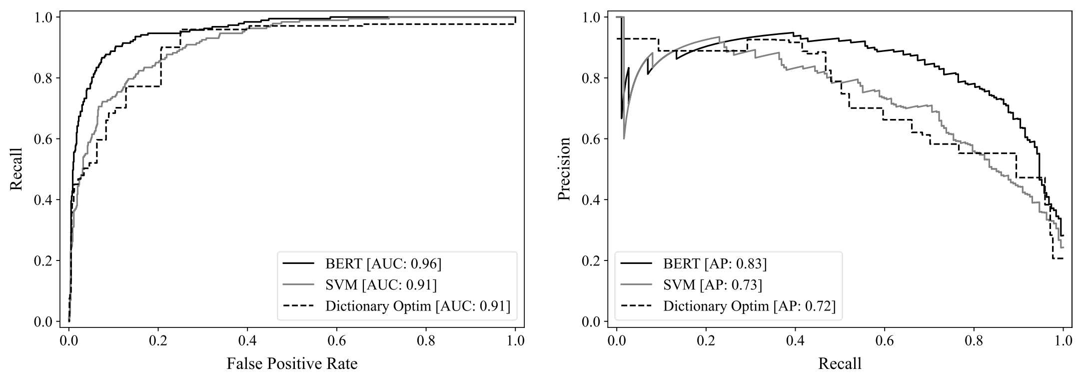

Figure 20. ROC and Precision-Recall Curves

To contextualize these performance measures, we compare the developed artifact with two alternative approaches. The first approach is the optimization of the dictionary. Instead of using a relatively unspecific but sensitive dictionary for preselection and increasing the precision through the machine learning module, only a dictionary is used, which is optimized based on the training and validation dataset (N=3,000) to increase its precision. For this purpose, we calculate for each word or n-gram of the dictionary how often it occurs in non-environmentally relevant sentences and how often it occurs in environmentally relevant sentences. The ratio between these two counts is calculated. A high ratio means, that the word or n-gram frequently occurs in non-environmentally-relevant sentences. These entries with a particularly high ratio could lead to many false-positives. Subsequently, the entries with the highest ratio (high occurrence in non-relevant sentences) are removed from the dictionary step by step. For evaluation, the sentences from the test dataset are classified based on this optimized dictionary. A sentence is considered environmentally relevant if it contains at least one of the entries. The performance of this benchmark is shown in Figure 20 by the dashed line (Dictionary Optim). To achieve the same recall as with the hybrid approach (0.8342), the precision must be reduced to 0.4722. This benchmark shows that applying a hybrid approach is indeed necessary and that it is not sufficient to reduce the entries in the dictionary until the precision has reached the desired value. As a second benchmark, we use a support vector machine trained on sentence-level averaged GloVe word embeddings (Pennington et al. 2014), represented by the grey line in Figure 20. It has been shown that the use of these word embeddings can improve supervised NLP systems (Pennington et al. 2014). The computational effort to train an SVM based on GloVe embeddings is significantly lower than fine-tune the BERT model on sentence classification. We chose the SVM because it is particularly suitable for high-dimensional and imbalanced classification problems (Goudjil et al. 2018), which is in line with our requirement (DR2). This benchmark is used to evaluate whether it is necessary to use such a complex model as BERT. The achievable precision is 0.5324 if a minimum value for the recall is set to 0.8342. The area under the curve (AUC) and the average precision (AP) show that the BERT model is superior to both benchmarks for this application.

After it has been shown that the artifact can extract environmentally relevant sentences from analyst reports with adequate recall and precision, we evaluate whether the extracted sentences are relevant for investors. For this purpose, we extract all environmentally relevant sentences from the corpus utilizing the artifact and identified 12,884 sentences as environmentally relevant. The preselection contains 65,848 sentences. Figure 21 shows a word cloud based on the whole corpus on the left-hand side and based on environmentally relevant sentences on the right-hand side. This gives a first impression of the content discussed in the extracted sentences.

Figure 21. Word Cloud from the Full Corpus and the Extracted Sentences

To analyze the topic structure of these sentences in more detail, we apply Latent Dirichlet Allocation (LDA) and develop a topic model (Blei et al. 2003). This methodology allows us to find topics that are discussed across different documents. Based on the coherence score and a visual inspection of varying topic models, we set the number of topics to 10, which is the central hyperparameter in LDA. Each sentence is considered to be a separate document. Table 8 shows the 20 most important words per topic. Based on these top words, we have assigned a label to each topic that describes it.

Topic: Label	Top 20 Words per Topic
1: CO_2 target	target, company, reduction, carbon dioxide emission, management, give, additional, current, achieve, follow, performance, meet, require, recent, line, reach, policy, regard, deliver, highlight
2: Expansion of renewables	increase, year, capacity, total, offshore wind, order, estimate, asset, addition, add, mw, unit, offshore, revenue, share, expect, turbine, strong, portfolio, farm
3: Efficiency of products	emission, cost, reduce, sale, level, change, carbon dioxide, term, start, fuel, assume, test, challenge, cut, show, standard, ahead, close, fleet, peer
4: Fuel and emission market	market, high, price, continue, demand, due, carbon, low, rise, remain, drive, decline, coal, offset, mix, margin, positive, trend, volume, expect
5: Regulatory	risk, result, potential, lead, view, regulation, future, key, time, benefit, opportunity, sector, industry, average, return, government, move, point, exist, company
6: Legal issues	impact, car, model, relate, include, issue, sell, fine, environmental, number, major, hybrid, provision, range, face, state, pay, launch, concern, brand
7: Clean products and production	technology, product, production, improve, customer, solution, waste, build, efficiency, offer, process, source, great, good, water, work, consumer, building, produce, aim
8: Diesel scandal	month, vehicle, diesel, make, report, system, european, engine, software, limit, announce, accord, case, german, control, full, investigation, measure, fix, today
9: Power generation	wind, power, energy, solar, project, gas, large, global, electricity, service, plant, supply, base, operation, generation, country, engie, construction, world, power generation
10: Power grid investments	renewable, growth, business, investment, network, focus, plan, grow, strategy, capex, grid, generation, area, segment, activity, period, exposure, group, main, represent

Table 8. Top Words of Topics from Extracted Sentences

In addition, we provide for each of the ten topics an example sentence in Table 9. The sentences have been chosen based on the probability for the respective topic assigned by LDA. Overall, it can be seen that CO_2 emissions and the resulting climate change play an important role in the extracted text sections. Topic 1 deals with it directly. Jonson et al. (2019) discuss the CO_2 reduction targets of the analyzed company (see topic 1 in Table 9). Topics 2, 3 and 4 are also related indirectly to the problem of greenhouse gas emissions. Topic 2 focuses on specific projects to reduce emissions. These are not only investments in wind or solar farms of large energy providers. Ferry and Letzeler (2018) for example reported that a brewery shifted towards an own renewable energy supply (see topic 2 in Table 9). The third topic thematized the energy efficiency and emissions of products. This is particularly noticeable in analyst reports on automotive manufacturers. This reveals another important characteristic of the environmental relevance of information for investors. Ellinghorst et al. (2019) link the environmental aspects with financial indicators (see topic 3). This link is perceived as very important by investors (IIRC & Kirchhoff 2020). Topics 5 and 6 deal with environmental regulation and legal disputes. Aguilar (2018) discusses a court settlement over a penalty for pollution. Again, this shows the linkage of environmental and financial issues. Topic 7 looks at transformation steps towards more sustainable production and product design. The focus is not only on CO_2 emissions but also on waste reduction and water consumption. Vasilescu et al. (2018) describe in their analyst report a production concept that reduces waste. Topic 8 summarizes statements related to the VW emissions scandal. As this had significant consequences for VW and was perceived worldwide, it is discussed extensively in many analyst reports. Topic 9 contains information about energy generation and its sources while topic 10 deals with grid investment projects. Here, for example, Mackie (2015) points out in his analyst report that growth opportunities for smart grids arise for the analyzed company due to the growing share of renewable energy.

Overall, it can be seen that the developed artifact can be useful for extracting targeted information on sustainability. These text fragments cover a broad range of environmental issues and link sustainability with financial indicators and are therefore highly relevant for investors. The information extracted by the artifact can provide an additional source of information to the quantitative ESG ratings and the self-reported information from CSR reports.

Topic	**Example Sentences per Topic**
1	*“On track to achieve its current target of a 25% reduction in carbon by 2020, in the coming months CRH will update the market on its ESG objectives putting out targets for both 2025 and 2030.”* (Jonson et al. 2019, p. 1)
2	*“Renewable energy: ABInBev has finalised a purchase agreement for on-site solar equipment at five South African breweries, which equates to 10% of South African electricity requirements*.” (Ferry and Letzeler 2018, p. 34)
3	*“Based on the industry’s 2018 CO2 footprint, we estimate it will cost an aggregate €15.1bn to comply, assuming a €60 cost per gram to reduce CO2 emissions for premium carmakers and €40 for volume OEMs (our discussions with companies suggest that this is a reasonable rule of thumb).”* (Ellinghorst et al. 2019, p. 14)
4	*“Mining is undergoing significant distress due to coal’s declining cost-competitiveness relative to unconventional natural gas production and more carbon emissions legislation.”* (Schoonmaker 2018, p. 9)
5	*“There are four ways in which the new President could benefit the industry: […] 4) new measures to improve energy efficiency could benefit Saint-Gobain.”* (Gardiner 2017, p. 2)
6	*“There was also the February legal settlement with the state of Minnesota that amounted to a pretax charge of $897 million, inclusive of legal fees and other obligations, related to its natural resource damages lawsuit concerning certain perfluorocarbons present in the environment.”* (Aguilar 2018, p. 15)
7	*“To reduce waste, much of Nike’s focus over the past few years has been on additive versus deductive manufacturing.”* (Vasilescu et al. 2018, p. 6)
8	*“The agencies accused the company of deliberately manipulating through software algorithms in roughly 428,000 diesel-equipped vehicles, the activation of anti-pollution controls during emissions tests only.”* (Hilgert 2015, p. 17)
9	*“Last year, in Texas >70m MWh of power was generated from renewable sources, enough to power 2mln homes for an entire year, and today AL announced it has signed an agreement for the supply of 50MW of renewable power, which will come on line at the end of 2020.”* (Walsh et al. 2018, p. 2)
10	*“In contrast, rising penetration of renewable energy generation in North America, Europe and Asia has supported growth and returns for Medium-Voltage products, Energy Automation and Smart Grid Solutions and Services, segments where we see 4% compound growth potential.”* (Mackie 2015, p. 15)

E.

Table 9. Per-Topic Examples of Environmental Sentences

6 Discussion

Our study shows that the proposed design principles and design features are suitable for extracting sustainability-relevant information from analyst reports. This contributes to solving the lack of meaningful and reliable information on corporate sustainability. The hybrid approach that combines a dictionary with a state-of-the-art machine learning model is a central solution component of the developed artifact. The dictionary used for preselection ensures a high level of sensitivity and at the same time a grounding on existing knowledge and generally accepted definitions of sustainability. Further, due to the extreme class imbalance, the preselection allows a comprehensive training and test dataset to be manually labeled in a reasonable amount of time. Through the machine learning component, a highly specific extraction is achieved. By taking the context into account, the model can deal with the frequent ambiguity of terms used in the sustainability and finance context.

Through a comprehensive evaluation of the classification performance and the content analysis of extracted sentences, it becomes apparent that the artifact can help investors to include sustainability aspects in their investment decisions. Because it provides investors with concrete and qualitative information, it supplements the often used and less transparent ESG scores. Unlike CSR reports, the extracted information is not self-reported and should therefore provide a closer look at critical aspects.

However, our proposed solution is also subject to some limitations. Only environmentally relevant topics were extracted during prototype development and evaluation. The other two dimensions of sustainability following GRI *economic* (GRI 200) and *social* (GRI 400) are not considered. The design principles and features developed should also be transferable to these, but a related replication has to be carried out first, as little can be said about the scope of this information in analyst reports. Thereby, future related research can contribute to more complete and mature knowledge and nascent design theory (Gregor and Hevner 2013).

Another limitation is the evaluation of sensitivity. Based on the sample, we estimate that 0.3% ($CI_{0.95}$ = [0.11%; 0.87%]) of the records contain sustainability information. However, the confidence interval is relatively large. If the proportion of sustainability sentences is actually 0.87%, many sustainability-relevant sentences have been overlooked during the preselection and the sensitivity of the artifact would be very low. Due to a high degree of uncertainty about the underlying proportion of environmental sentences in the total population, a reliable statement about the sensitivity can only be made starting from the preselection.

The evaluation is also not yet exhaustive. Despite the comprehensive content analysis of the extracted sentences, an evaluation involving investment practitioners has to be carried out in order to assess the decision usefulness of the information obtained in a real-world scenario.

7 Conclusion

In this paper, we present a problem solution and related design principles for extracting relevant information on sustainability from analyst reports, which are available to many investors. In doing so, we address the obstacles that prevent investors from fully integrating sustainability considerations into their investment strategy. These obstacles are, among others, the lack of data, the costs of information gathering and the non-specificity of self-disclosed information on sustainability (Amel-Zadeh and Serafeim 2018; PwC 2020). Building on a hybrid architecture combining a knowledge grounded preselection and a state-of-the-art machine learning model, the artifact can improve investors' information base. This should improve the decisions of investors concerning the sustainability assessment of companies. Better informed decisions will improve the allocation of capital, shifting real investments from non-sustainable to sustainable companies. As a result, it can contribute to a more sustainable economy and thus to the achievement of the SDGs.

We contribute to the literature by proposing design principles and features that address the specific problems of class imbalance and the need for a unified understanding of sustainability. The generalizability of our approach, however, allows theorists and practitioners alike to apply this approach to related problems. Especially against the background of ever-increasing data volume, information extraction will become more and more important. In addition to its use in investment practice, our artifact might be helpful for researchers conducting research on corporate sustainability or SRI. They can apply the artifact to extract sustainability-relevant information from financial documents.

Our work offers starting points for subsequent design science research. First, the artifact should be extended with respect to the dimensions *economy* and *social*. Second, a field study with investors should evaluate the extent to which the extracted information is useful for decision-making. Third, the functionality of the artifact should be further developed and corresponding design principles and features derived. There is particular potential for improvement in the presentation of the extracted sentences. One extension that should be discussed with potential users would be to display a snippet from the PDF document for each extracted sentence so that the investor can read the sentence in context and supplementing graphics and tables can thus also be considered.

II.2. MiFID II and Analyst Reports

MiFID II and its Effects on Analyst Reports

Authors Steffen Bankamp, Jan Muntermann and Matthias Palmer

Outlet Working paper, prior version accepted and presented at the INQUIRE Autumn Residential Seminar, Bath, United Kingdom, 2021.

Abstract: In early 2018, the European regulation “Markets in Financial Instruments Directive II” came into force that especially affected sell-side analysts and asset managers. One central aspect of this regulation is the so-called research unbundling, which requires fees for execution and research services to be separately charged by brokers. Our paper investigates how this regulatory change alters company coverage, optimism, and the novelty of textual content conveyed in analyst reports. The results reveal less optimism and higher novelty, indicating an increase in quality, while we observe reduced company coverage, indicating a decrease in quantity. Most importantly, two well-documented behavioral biases are reduced: analyst optimism and herd behavior.

Keywords: MiFID II, regulation, financial analysts, equity research, research unbundling

1 Introduction

In January 2018, the European Directive 2014/65/EU (EU 2014), better known as "Markets in Financial Instruments Directive II" (MiFID II), came into effect. MiFID II aims to harmonize the European financial market, increase transparency, and protect investors. This regulation has an extensive impact on financial analysts, as it changes their business model. The "research unbundling" component of the regulation requires the execution of orders and research services offered by brokers to be invoiced separately ("hard dollars"), eliminating the usual practice of brokers' research being cross-subsidized via execution services ("soft dollars"). As a result, asset managers must decide whether they bear the research costs themselves (offsetting against their profit and loss statement or pass the costs on to their investors via research budgets. The overall goal of "research unbundling" is to increase cost transparency for investors.

Textual analyst reports are a key element in financial analysts' work (A.H. Huang et al. 2018). In these reports, sell-side analysts provide both forecasts and detailed qualitative explanations. The latter discuss current company news, assess the company in the overall market, and interpret managements' statements from conference calls. While forecasts are the quintessence of analysts' opinions, a forecast without sufficient explanations is of limited use for a buy-side asset manager. With "research unbundling" taking effect, a limited budget, and the need for an asset manager to decide which research provider to use, the quality of an analyst report is likely to be a major determinant. It may even be more important than the quality of quantitative analyst estimates, as they are rarely considered individually but rather as a consensus across different analysts. Given these developments, asset managers might apply higher standards, especially to the quality of the textual analyst reports, than was the case before the explicit pricing of analyst research services.

With a limited amount of time and resources available to financial analysts, they must set priorities to achieve quality leadership in their chosen sectors or companies and market their reports successfully. Furthermore, financial analysts employed by brokers are said to publish very optimistic recommendations, as they are incentivized by the trading commissions generated from their recommendations (Beyer and Guttman 2011; Cowen et al. 2006; Irvine 2004; Jackson 2005). Under short sales constraints, analysts can generate larger volumes from buy recommendations than from sell recommendations, as the trading volume for sell orders would be restricted to the investor's initial stock holdings (Hayes 1998). Irvine (2004) also empirically shows that analysts can generate higher trading volumes from buy than from sell recommendations. As MiFID II detaches the research departments from the fees generated via execution services, this should reduce the incentive for sell-side analysts to produce over-optimistic analyses. At the same time, the explicit pricing should lead analysts to create more differentiated analyses and thus justify the price charged to investors.

One approach in adding further economic value could be for analysts to integrate additional non-public information into their reports to differentiate themselves from peers. In the past, the incentives, especially for young analysts, were exactly the opposite. That is, herding has been reasonable, resulting in analysts following the consensus and paying less attention to private information (Hong et al. 2000). With the introduction of “research unbundling,” we assume that when asset management evaluates the reports in more detail during price negotiations, it can quickly become visible if entire text passages or lines of argumentation are replicated.

Changes brought about by MiFID II’s “research unbundling” are intended to ensure that decisions about acquiring sell-side research are made in the interests of the investors, who ultimately pay for it. In this paper, we thus highlight the impact of MiFID II on analyst reports by leveraging state-of-the-art text mining techniques. The results reveal a decrease in analyst coverage and show that the introduction of MiFID II has reduced analyst optimism. Furthermore, sell-side analysts provide additional information in their reports and repeat less content of their peers that is already known.

Our study contributes to the literature stream on research unbundling in general (see the literature review of Bender et al. (2021)) and to the effects of MiFID II on financial analysts in particular (see Amzallag et al. 2021; Anselmi and Petrella 2021; B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021; Z. Liu and Yezegel 2020; Pope et al. 2019; Preece 2017, 2019). These early studies have found that analysts stop covering especially less lucrative companies – they react to the pressure created by the new pricing model. Prior research has also found more accurate earnings per share forecasts, indicating an improvement in quality (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021). Although prior research has thoroughly studied the impact of MiFID II on the quantitative part of analyst output (e.g., forecasts and recommendations), the extent to which MiFID II has influenced analyst reports and thus qualitative analyst output is completely unknown. We agree with A.H. Huang et al. (2014a), who state that a holistic view of financial analysts’ work can only be considered when qualitative analyst output is also considered.

From a general perspective, our study contributes to the literature on behavioral biases among financial analysts (e.g., Francis 1997; Hong et al. 2000; Mehran and Stulz 2007), and we support answering the question of how financial regulation can mitigate these biases (e.g., Barber et al. 2006; Bradshaw et al. 2019; Espahbodi et al. 2015).

The practical relevance of our study is confirmed by the three studies from the Chartered Financial Analyst (CFA) Institute on the effects of MiFID II (Allen 2019; Preece

2017, 2019) that observe reduced research budgets and increased competition, and a position paper proposing regulatory actions similar to those in MiFID II in the US jurisdiction (Allen and Gellasch 2019). Overall, our research aims to support asset managers in assessing overall analyst report quality, provides implications for regulators, and reveals the intended and unintended consequences of "research unbundling" as part of MiFID II.

2 MiFID II, Research Services, and Cost Transparency

MiFID II (Directive 2014/65/EU; EU (2014)) is a European regulation that came into force on January 3, 2018. It succeeds the MiFID regulation (Directive 2004/39/EC; EU (2004)) implemented in 2007. Both MiFID and MiFID II aim for the harmonization of the European financial market. The regulation intends to create an internal market in which investment firms can offer their services across the EU. Transparency should be increased and investors protected by the uniform regulation of the European financial market. MiFID II extends the regulations from MiFID and considers developments in the financial market and experiences from the financial crisis. The MiFID II regulation results in fundamental changes for "investment firms, market operators, data reporting services providers, and third-country firms providing investment services or performing investment activities through the establishment of a branch in the Union" (Directive 2014/65/EU Article 1(1); EU (2014)). A major aim of MiFID II is to improve the protection of investors. Therefore, it covers various regulatory changes (e.g., algorithmic trading, multilateral trading facilities, and best execution). One such case is the "research unbundling" introduced in the motivation of this paper.

The location of the investment company determines whether MiFID II must be applied. This regulation is mandatory for all investment companies domiciled in a member state of the European Economic Area (EEA). The EEA consists of all members of the EU, Iceland, Liechtenstein, and Norway. The United Kingdom was bound to EEA rules and regulations and thus to MiFID II until the end of the transmission period on December 31, 2020. Asset managers who must comply with the MiFID II regulation due to their location must insist that their brokers show the costs separately. This also explains why MiFID II's influence cannot be limited to the EEA alone, as brokers from the US, for example, must also bill the amount for execution and research services for their European clients separately.

3 The Impact of MiFID II on Analyst Behavior

In anticipation of the introduction of MiFID II, the CFA Institute (Preece 2017) surveyed investment professionals working for the buy side on the expected effects of MiFID II in 2017. Participants had a clear expectation that the buy side would acquire less research and that they would charge the expenditure for analyst research to the profit and loss statement of the buy side and not pass it on directly to investors. Based on a decline in buy-side demand, it is unclear how this affects the sell side.

Prior research has demonstrated an analyst coverage loss that is particularly noticeable for companies with low importance for the sell side (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021; Z. Liu and Yezegel 2020). B. Fang et al. (2020) find that these firms are small, have low institutional ownership, and currently do not issue equity or debt. According to Lang et al. (2021), companies will lose coverage if an additional analyst has low incremental value. However, the results of Lang et al. (2021) are not entirely consistent with those of B. Fang et al. (2020). They find more substantial coverage losses for companies that are large and old, have low volatility, were previously very strongly covered, and for which the analysts' consensus forecast was relatively accurate. Guo and Mota (2021), Amzallag et al. (2021), and Anselmi and Petrella (2021) find that large firms experience substantial coverage losses, whereas small firms' coverage remains stable. B. Fang et al. (2020) conclude that the buy side substitutes the sell-side coverage loss by employing more in-house analysts. In another survey of the CFA Institute, Allen (2019) found that MiFID II not only affects Europe but also has an impact on the investment research market globally. Hence, 42% of the surveyed investment professionals from the US rate the market for investment as being more competitive than prior to MiFID II. In comparison, only 10% find the market to be less competitive.

Aside from analyzing the number of analyst forecasts, previous research has also investigated analysts' forecast quality. There is strong empirical evidence for an increase in non-aggregated earnings-per-share (EPS) forecast accuracy (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021). However, in a difference-in-differences setting, comparing broker and non-broker analysts, Z. Liu and Yezegel (2020) find that the forecast accuracy of small brokers decreases. Furthermore, in a survey of buy-side and sell-side investment professionals, Preece (2019) reports that 44% of the sell-side and 27% of the buy-side professionals believe that research quality has decreased, while only 9% and 4% of the respondents, respectively, think that research quality has increased.

Given these major findings on quantity and quality, the studies provide further evidence regarding various aspects. B. Fang et al. (2020) and Lang et al. (2021) find

MiFID II to increase bid-ask spreads and thus to lower market liquidity. Moreover, Lang et al. (2021) observe a delay in the information provision of analysts. Individual forecasts and stock recommendations lead to stronger market reaction compared to the pre-MiFID time frame (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021; Z. Liu and Yezegel 2020). However, on the aggregated level, the market reaction decreases due to MiFID II (Guo and Mota 2021; Lang et al. 2021). Guo and Mota (2021) argue that the difference is due to the opposing effects of the regulation. The decrease in the quantity of analyst output deteriorates market efficiency, while the increase in quality improves it. The reduced market reaction on the aggregate level indicates a quantity effect that outweighs the quality effect. However, Anselmi and Petrella (2021) could not find a negative impact of MiFID II on stock market liquidity and price efficiency.

The aforementioned studies provide initial evidence of the effects of MiFID II by relying on forecasts and recommendations from I/B/E/S. Only Z. Liu and Yezegel (2020) consider the number of pages of analyst reports and, therefore, a report-specific measure. Thus, with our text-mining approach, we capture a previously neglected perspective on the work of financial analysts in light of MiFID II. The surveys conducted by the CFA Institute (Allen 2019; Preece 2017, 2019) and the results of the aforementioned studies indicate that MiFID II has increased the pressure on sell-side analysts, who must now justify to their clients that investment research services are worth the price they charge. This raises the question of how analysts modify their reports to respond to the changed market conditions induced by MiFID II. To answer this research question, we examine analyst reports based on three dimensions: coverage, analyst optimism, and novelty. Coverage focuses on quantity, whereas analyst optimism and novelty are two aspects of analyst reports' quality.

Before analyzing the reports' content in detail, we evaluate whether the quantity effect observed on I/B/E/S data also applies to analyst reports. Not every I/B/E/S entry is followed by an analyst report, and not all analysts who publish reports are registered on I/B/E/S. By replicating previous research on coverage, we ensure that the dataset we use is suitable for determining the effects of MiFID II.

Theory on product bundling suggests that it can be used to skim consumer surplus (Schmalensee 1984). In a competing market, product bundling is a form of product differentiation to reduce competition (Y. Chen 1997). As the ability to bundle products is eliminated, competition among analysts should increase. This theoretical consideration is in line with the empirical observations from Preece (2019), who finds a more competitive research marketplace. Unlike before, customers can now assess the value of execution and research services separately, and asset managers will make their procurement decision based on this evaluation. Analysts who cannot explain the added value of their research services to their clients will not generate sufficient revenue from selling their reports to cover the costs of conducting this research and thus discontinue the coverage. This should lead to a decreasing number of analysts publishing reports after the implementation of MiFID II.

It can be assumed that transforming the research department of brokers from a cost to a profit center reduces the dependency on trading commission and thus reduces conflicts of interest (B. Fang et al. 2020). To a certain extent, these conflicts of interest have led to analyst optimism (Jackson 2005). In addition, with the increased transparency and competition among analysts, they might differentiate themselves from competitors by providing more realistic analysis. Thus, we expect that this optimism bias will be reduced and that the textual content in analyst reports will be written less optimistically and will provide a more realistic company assessment.

Existing literature shows stronger market reactions following individual analyst forecasts and recommendations. This indicates an increase in the informativeness of analyst output (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021; Z. Liu and Yezegel 2020). However, it is unclear how analysts increase the informativeness of their research output. We argue that analysts are increasing their reports' informativeness by incorporating more private information into the reports and relying less on public information that has already been published by peers. This is the opposite case to herding behavior, which has been intensively observed in the context of financial analysts (e.g., Trueman 1994; Hong et al. 2000). The explicit pricing model required by MiFID II will lead to increased pressure on the sell side. Analysts must more precisely explain the added value of their analyses to customers (B. Fang et al. 2020); this can be achieved by including more relevant and private information in the reports. Since analysts might supposedly start to put more weight on their private information after the MiFID II implementation, reports should differ more clearly from one another and contain further information that was not previously included in other analysts' reports.

4 Methodology

To evaluate the effects of MiFID II on analyst reports, we use a difference-in-differences approach. This approach is particularly suitable for evaluating a treatment that applies to large groups simultaneously (e.g., policy changes in an entire country) and therefore does not have a variation of the treatment assignment within this group (Angrist and Pischke 2008).

Financial analysts are confronted with multiple changes that have occurred in recent years, for example competition through FinTechs and the use of artificial intelligence (Grennan and Michaely 2021, 2020). Thus, analyst output before and after regulation cannot be simply compared, as changes may be due to not only the introduction of MiFID II but also other changes in analysts' business environment. The difference-in-differences approach allows us to control for these trends if they are global and therefore affect the treatment as well as the control group. The difference-in-differences approach has already been employed in other studies on MiFID II (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021). Since MiFID II is a European regulation, we use European companies as a treatment group and US companies as a control group. This setting is also utilized in the abovementioned studies. However, the applied separation of control and treatment group is not ideal, as the jurisdiction of MiFID II is linked to the location of asset managers and not to the location of the company that is covered. Since the location of asset managers who buy investment research is unknown, we must rely on this heuristic. Guo and Mota (2021) argue that this approach is reasonable due to the home bias of investors (see e.g., Coval and Moskowitz 1999; Feng and Seasholes 2004). When making investment decisions, the preference for local firms should also result in asset managers buying relatively more analyst reports covering companies located in the same region as the asset manager. This results in financial analysts preparing to distribute more of their reports on European (US) companies to European (US) asset managers. However, it should be noted that this allocation of treatment and control groups is not entirely discriminatory. At the same time, it implies that the effects determined using this approach are rather under- than overestimated.

4.1 Coverage

To examine the impact of MiFID II on analyst coverage of I/B/E/S and analyst reports, we estimate Equation (1):

$$\begin{aligned}
&\mathrm{COVEARAGE[IBES;REPORTS]}_{t,i} = \alpha_0 + \alpha_1 \mathrm{POST_TREAT}_{t,i} \\
&+\alpha_2 \mathrm{ACTIVE_FUNDS}_{t,i} + \alpha_3 \mathrm{MTB}_{t,i} + \alpha_4 \mathrm{DEBT_RATIO}_{t,i} + \alpha_5 \mathrm{SIZE}_{t,i} \\
&+\alpha_6 \mathrm{ROA}_{t,i} + \alpha_7 \mathrm{PPE_RATIO}_{t,i} + \alpha_8 \mathrm{LOSS}_{t,i} + \alpha_9 \mathrm{GDP_GROWTH}_{t,i} \\
&+\alpha_{10} \ln(\mathrm{GDP_CAPITA}_{t,i}) + \alpha_i + \alpha_t + \varepsilon_{t,i}
\end{aligned} \tag{1}$$

The observations are on a company-year level. For I/B/E/S, we measure the COVERAGE as the number of unique analysts submitting an EPS forecast on company i during year t to I/B/E/S. To measure COVERAGE based on analyst reports, we follow the same approach and count the unique analysts that published an analyst report on company i during year t. Of particular interest is the interaction term POST_TREAT, which takes a value of 1 for all observations that took place after the introduction of MiFID II (January 3, 2018) and relate to the treatment group (European companies). For the remaining observations, the interaction term takes a value of 0. Based on the discussed theory and prior studies, we expect a significant negative regression coefficient for α_1. We follow existing literature and include company and country-specific control variables that could influence analyst coverage (B. Fang et al. 2020). These control variables include MTB (market to book ratio), DEBT_RATIO, SIZE (measured as the natural logarithm of total assets' United Sates Dollar [USD] value), ROA (return on assets), PPE_RATIO (book value of property plant and equipment relative to total assets), LOSS (a variable that becomes 1 if the company generated a loss in the respective year), GDP_GROWTH, and ln(GDP_CAPITA) (natural logarithm of the gross domestic product [GDP] per capita measured in USD). In addition, we control for the ratio of active to passive fund ownership (ACTIVE_FUNDS). Since analyst reports should only be relevant to investors with an active investment approach, this variable allows us to control for a heterogeneous shift in ownership between active and passive funds across firms that potentially influences the demand of analyst reports. A detailed variable definition can be found in Table 17. To control for all time-invariant company characteristics and all period-specific characteristics, we include firm fixed effects α_i and year fixed effects α_t, respectively.

4.2 Optimism

To understand how MiFID II has affected analyst optimism, we examine optimism on the individual report level indicated by the index j. We estimate Equation (2):

$$\begin{aligned}
&\mathrm{SENT_POLARITY}_{j,t,i} = \alpha_0 + \alpha_1 \mathrm{POST_TREAT}_{t,i} + \alpha_2 \mathrm{ACTIVE_FUNDS}_{t,i} \\
&+\alpha_3 \mathrm{MTB}_{t,i} + \alpha_4 \mathrm{DEBT_RATIO}_{t,i} + \alpha_5 \mathrm{SIZE}_{t,i} + \alpha_6 \mathrm{ROA}_{t,i} + \alpha_7 \mathrm{PPE_RATIO}_{t,i} \\
&+\alpha_8 \mathrm{LOSS}_{t,i} + \alpha_9 \mathrm{GDP_GROWTH}_{t,i} + \alpha_{10} \ln(\mathrm{GDP_CAPITA}_{i,t}) \\
&+\alpha_{11} \mathrm{RETURN}_{j,t,i} + \alpha_{12} \mathrm{VOLATILITY}_{j,t,i} + \alpha_k + \alpha_i + \alpha_t + \varepsilon_{j,t,i}
\end{aligned} \tag{2}$$

As a dependent variable, we use the sentiment polarity (SENT_POLARITY) extracted from the reports' text. To do so, we follow a machine learning approach and count the positive and negative classified sentences that occur within the analyst reports and transform these two values into a polarity measure by applying Formula (3) (Henry 2008; Uang et al. 2006):

$$\mathrm{SENT\_POLARITY} = \frac{\mathrm{N_{pos}} - \mathrm{N_{neg}}}{\mathrm{N_{pos}} + \mathrm{N_{neg}}} \tag{3}$$

The polarity contains a value of 1 (−1) if the text entirely consists of positive (negative) and neutral sentences. A high polarity thus indicates an overly optimistic written analyst report. To identify positive, negative, and neutral sentences, we use a classifier called FinBERT, which was trained and evaluated by Y. Yang et al. (2020). This classifier is based on a state-of-the-art language representation model from Devlin et al. (2019). Y. Yang et al. (2020) showed on a sample of analyst reports that this classifier has an accuracy of 88.7% in correctly predicting a sentence's sentiment. It is, therefore, well suited for our application, as it achieves higher accuracy than sentiment classifiers based on Naïve Bayes or dictionaries (A.H. Huang et al. 2014a). As in the coverage analysis, we are interested in the interaction term POST_TREAT. To avoid indirectly observing an effect of coverage loss (e.g., optimistic analysts dropping out), we follow B. Fang et al. (2020) and use a balanced sample on the broker–company level. Only reports from company–broker combinations are used, which are available in the dataset both before and after introducing MiFID II. Aside from the control variables and fixed effects of the previous analysis, broker fixed effects α_k are included. Furthermore, we control for the stock return of the 60 trading days before the publication of the report (RETURN) to avoid our results being driven by a different performance of European and US companies.

4.3 Novelty

To evaluate how analysts alter the informativeness of their reports after the MiFID II introduction, we measure the amount of new information contained in a report. This is operationalized by the NOVELTY_SCORE, which stems from the research field of novelty detection and has already been shown to be suitable to detect documents with novel content (Yi Zhang et al. 2002). Brown and Tucker (2011) applied a similar approach to evaluate year-over-year modification of the MD&A section in annual reports. The NOVELTY_SCORE is appropriate for determining whether a particular analyst report provides new information compared to a set of older reports. We estimate Equation (4) to determine MiFID II's impact on analyst reports' novelty:

$$\begin{aligned}
&\mathrm{NOVELTY_SCORE[DAYS]}_{j,t,i} = \alpha_0 + \alpha_1 \mathrm{POST_TREAT}_{t,i} \\
&+\alpha_2 \mathrm{ACTIVE_FUNDS}_{t,i} + \alpha_3 \mathrm{MTB}_{t,i} + \alpha_4 \mathrm{DEBT_RATIO}_{t,i} \\
&+\alpha_5 \mathrm{SIZE}_{t,i} + \alpha_6 \mathrm{ROA}_{t,i} + \alpha_7 \mathrm{PPE_RATIO}_{t,i} + \alpha_8 \mathrm{LOSS}_{t,i} + \alpha_9 \mathrm{GDP_GROWTH}_{t,i} \\
&+\alpha_{10} \ln(\mathrm{GDP_CAPITA}_{t,i}) + \alpha_{11} \mathrm{RETURN}_{j,t,i} + \alpha_{12} \mathrm{VOLATILITY}_{j,t,i} \\
&+\alpha_k + \alpha_i + \alpha_t + \varepsilon_{j,t,i}
\end{aligned} \tag{4}$$

The structure of Equations (4) and (2) is identical. We also use a balanced sample at the broker–company level to avoid the coverage effect biasing the results. We choose the report level as the level of observation and the same control variables and fixed effects as in Equation (2).

Novelty is an essential quality feature of analyst reports. Information that is common knowledge is already included in the security's market price if the semi-strong form of market efficiency holds (Fama 1970). This means that there is no possibility of exploiting this information by implementing information-based trading strategies. On the other hand, asset managers should endeavor to exclusively secure those analyst reports that contain non-public information. We compare each analyst report with all analyst reports published by other brokers in a defined period before publication, and we calculate the mean cosine distance to these documents. We call this value NOVELTY_SCORE and assign it to each document. If a low novelty score is found, the analyst report is similar to previous reports; it thus contains ample already known information and consequently offers less value to its readers.

The NOVELTY_SCORE is defined as the mean distance of the report j to all reports published in a given period before the publication of report j. Reports of the same broker are not considered, as documents from the same publisher are often similar due to their identical structure or writing style (Krakow and Schäfer 2020). Our approach thus ensures that the results are not driven by these circumstances that would lead to an underestimation of the NOVELTY_SCORE. We use periods of 10 and 30 days. To ensure a clear separation between pre-treatment and post-treatment periods within the difference-in-differences setting, we exclude observations that take place during the first 10 or 30 days of the post-treatment period for the NOVELTY_SCORE[10/30]. This ensures that no pre-treatment reports are indirectly included in the novelty calculation of the post-treatment reports. To measure the distance between two documents, we use the cosine distance of the two bag-of-words vectors. The word order within the document is irrelevant, as the measure is only influenced by the relative frequency of the words within the documents (Manning et al. 2008). The cosine distance is a measure for the angle between two word vectors in a high-dimensional space. If the two documents contain similar words in a similar relation, the angle, and hence the cosine distance, is small. We do not apply the cosine similarity to the standard word vector but to a vector weighted by its term frequency–inverse document frequency (TF-IDF). The TF-IDF weighting assigns high (low) weights to rarely (frequently) occurring words and thus prevents the analysis from being dominated exclusively by very frequent words. Furthermore, A.H. Huang et al. (2018) use the cosine similarity on bag-of-words vectors in the context of analyst reports. The approach is commonly used by researchers in the finance and accounting literature to compare text documents (Loughran and McDonald 2016).

In addition, we examine the novelty of analyst reports and thus analysts' herd behavior from a slightly different angle. A considerable proportion of analyst reports are published in close temporal relation to the earnings announcement (EA) date (A.H. Huang et al. 2018). EAs are important events for analysts. We use a time horizon from the day prior until the day after the EA and calculate the mean cosine distance between these reports if at least three reports from different brokers were published during this period. This procedure prevents the results from being driven by EAs, which only receive minimal reporting. In contrast to the previous analysis based on the NOVELTY_SCORE, the chronological order is not relevant for this analysis. For a better distinction, we subsequently call the variable COS[EARN_DATE]. With this approach, we evaluate whether analysts differentiate themselves more strongly if they publish analyst reports on a specific event. We calculate Equation (5) to evaluate the effect of MiFID II on COS[EARN_DATE]:

$$\begin{aligned}&\mathrm{COS[EARN\_DATE]}_{q,t,i} = \alpha_0 + \alpha_1 \mathrm{POST\_TREAT}_{t,i} + \alpha_2 \mathrm{ACTIVE\_FUNDS}_{t,i}\\&+\alpha_3 \mathrm{MTB}_{t,i} + \alpha_4 \mathrm{DEBT\_RATIO}_{t,i} + \alpha_5 \mathrm{SIZE}_{t,i} + \alpha_6 \mathrm{ROA}_{t,i} + \alpha_7 \mathrm{PPE\_RATIO}_{t,i}\\&+\alpha_8 \mathrm{LOSS}_{t,i} + \alpha_9 \mathrm{GDP\_GROWTH}_{t,i} + \alpha_{10} \ln(\mathrm{GDP\_CAPITA}_{t,i})\\&+\alpha_{11} \mathrm{RETURN}_{q,t,i} + \alpha_{12} \mathrm{VOLATILITY}_{q,t,i} + \alpha_i + \alpha_t + \varepsilon_{q,t,i}\end{aligned} \quad (5)$$

We apply the difference-in-differences approach to identify the effects of MiFID II at the quarterly firm-earnings level q. We expect that the reports published in this short time frame will be more distinct from one another, as analysts have less incentive for herding behavior under the MiFID II regulation.

5 Dataset

We obtain analyst reports from Refinitiv Thomson ONE. The sample is based on all companies that have been a constituent of the EURO STOXX 50 (treatment group, $N = 57$) as well as in the Dow Jones Industrial Average (control group, $N = 31$) at any point in time during the observation period ranging from 01-01-2015 to 12-31-2019. We exclude DuPont and its successors (DowDuPont and Dow Inc.) from our analysis to prevent structural breaks (a merger between DuPont and Dow Chemicals in 2017 and later spin-off of DuPont in 2019) affecting our analysis.

The selection of major US and European equity indices ensures sufficient coverage of the companies by analysts, which allows us to compare the analyst reports in a meaningful way. Following B. Fang et al. (2020), we choose a pre-treatment period of three years. This design allows us to access the parallel trend assumption required by the difference-in-difference approach (Egami and Yamauchi 2021). We further observe two years during the post-treatment period, making our sample more balanced than a setting that only utilizes the first year after the treatment occurred (Z. Liu and Yezegel 2020). This results in a total of 85,081 analyst reports. Pricing data and fundamentals used as control variables in the subsequent analysis are from Refinitiv Datastream, Refinitiv Eikon, and World Bank. To increase data quality, we remove reports not written in the English language, automatically generated analyst reports, the report with the most recent timestamp of duplicates, as well as short reports containing less than 300 words or exceptionally long reports with more than 50 pages. This results in a final sample of 60,626 reports from 209 unique research providers. The technical pre-processing steps of the analyst reports are explained in Appendix B. Of the 209 research providers, 158 offer broker services, while 51 are non-broker research providers. The 158 brokers account for 51,955 (85.70%) analyst reports. For the sake of simplicity, we refer to all research providers as brokers in the following.

	European Brokers	**US Brokers**	**Other Brokers**
European Firms	20,701	3,366	4,171
US Firms	13,583	14,993	3,812

Table 10. Contingency Table Between Broker and Firm Location

As can be derived from Table 10, a strong relationship exists between broker and company location. This relationship is especially strong for European companies, since 73.31% of the reports covering these companies are published by European brokers. However, European brokers only account for 56.55% of our sample's entire reports. This is also in line with Guo and Mota (2021) findings on analysts' strong geographical focus. For descriptive statistics of the dependent, independent, and control variables, see Table 11, which lists all variables by the utilized level of observation.

Variables	***N***	**Mean**	***SD***	**P25**	**Median**	**P75**
Firm-year level						
COVERAGE[I/B/E/S]	440	27.3080	5.9396	15.0000	24.0000	27.0000
COVERAGE[REPORTS]	440	20.0136	4.7746	11.0000	17.0000	20.0000
ACTIVE_FUNDS	440	0.6411	0.1134	0.3685	0.5659	0.6595
MTB	440	5.5398	36.7346	21.6399	1.2475	2.3400
DEBT_RATIO	440	0.4759	0.2270	0.0053	0.3184	0.4462
SIZE	440	25.5661	1.2321	23.2151	24.6208	25.4556
ROA	440	0.0585	0.0519	−0.0288	0.0205	0.0480
PPE_RATIO	440	0.2072	0.1966	0.0016	0.0592	0.1400
LOSS	440	0.0546	0.2274	0.0000	0.0000	0.0000
GDP_GROWTH	440	0.0174	0.0173	0.0029	0.0089	0.0157
ln(GDP_CAPITA)	440	10.7562	0.2476	10.1555	10.6089	10.7420
Earnings-announcement level						
COS[EARN_DATE]	1,658	0.6534	0.1067	0.3946	0.5753	0.6596
Report level						
NOVELTY_SCORE[10]	52,884	0.6745	0.1561	0.5664	0.6889	0.7932
NOVELTY_SCORE[30]	58,687	0.6845	0.1451	0.5831	0.6991	0.7957
SENT_POLARITY	60,626	0.2725	0.4267	0.0000	0.3333	0.5955
RETURN	60,626	0.0208	0.1099	−0.0405	0.0272	0.0907
VOLATILITY	60,626	0.2266	0.0837	0.1672	0.2126	0.2703

Table 11. Full Sample Summary Statistics

6 Results

6.1 Coverage

Figure 22 depicts the average yearly coverage of the companies based on I/B/E/S data (left plot) as well as on analyst reports from Refinitiv Thomson ONE (right plot). The coverage is measured as the number of unique analysts following the respective company during a calendar year. The bar plot shows the change in the difference between the European and US companies. Based on I/B/E/S data, for US firms we see a very stable coverage of approximately 28 analysts per company over time. For the European sample, a decline in coverage can be observed from 2015 to 2018. The most substantial decline is observable in 2018, when the MiFID II regulation came into place. In 2019 this effect reversed, and the coverage of European companies has increased again.

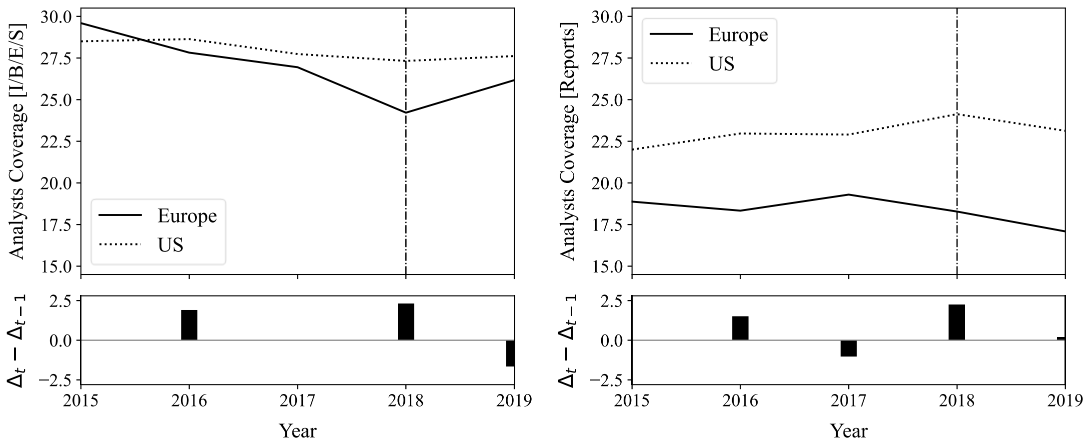

Figure 22. *Changes in Analyst Following Over Time, Derived From I/B/E/S and Textual Analyst Reports*

When examining textual analyst reports on the right-hand plot, the overall coverage is visibly lower. This is because not every price target or forecast stored within I/B/E/S has to be accompanied by an analyst report, which in turn is stored in the Refinitiv Thomson ONE database. However, not only the absolute values but also the development over time is different. The coverage of US companies increases continuously from 2015 to 2018 and only decreases in 2019. For European companies, a decrease in coverage is particularly noticeable in 2018 and 2019.

For both I/B/E/S data and textual analyst reports, we obtain a significant negative regression coefficient for the interaction term (POST_TREAT) (see Table 12). This indicates that the coverage of European companies compared to US companies has decreased after the introduction of MiFID II. Therefore, the result of lower analyst coverage is valid not only on forecasts from I/B/E/S but also on textual analyst reports. The European companies in our sample have lost, on average, 2.145 analysts compared to the control group. This is slightly more than the observed coverage reduction in I/B/E/S, which is, on average, 1.977 analysts. To account for heteroskedasticity and residuals that might be correlated across companies, we utilize robust standard errors clustered on the company level (Petersen 2009).

Variables	[1] COVERAGE [IBES]	[2] COVERAGE [REPORTS]
POST_TREAT	−1.977**	−2.145**
	(−3.62)	(−4.61)
ACTIVE_FUNDS	−2.319	3.698
	(−0.64)	(0.99)
MTB	0.004	0.003*
	(1.96)	(2.28)
DEBT_RATIO	6.433*	3.389
	(2.23)	(1.84)
SIZE	0.696	−0.062
	(0.81)	(−0.07)
ROA	−3.319	6.339
	(−0.60)	(1.11)
PPE_RATIO	−6.807	−3.394
	(−1.48)	(−0.72)
LOSS	−0.894	1.114*
	(−1.80)	(2.50)
GDP_GROWTH	−12.24	−8.361
	(−1.50)	(−1.73)
ln(GDP_CAPITA)	−8.689	−6.937
	(−1.12)	(−0.96)
INTERCEPT	103.8	93.19
	(1.29)	(1.19)
Observations	440	440
Fixed Effects	Firm & Year	Firm & Year
Clustering	Firm	Firm
Adj. R^2	0.862	0.831

** *p*-value < 0.01; * *p*-value < 0.05; *t*-values reported in parentheses

Table 12. Regression Results – Analyst Coverage Based on I/B/E/S Data and Textual Analyst Reports

To evaluate whether the impact of MiFID II is also of practical relevance, we calculate the effect size. Therefore we utilize Cohen's *d* (J. Cohen 1988). This measure relates the magnitude of our variable of interest (POST_TREAT) to the standard deviation of the dependent variable. To calculate the standard deviation, we follow Morris (2008) and use the pooled standard deviation of the treatment and control groups during the pre-treatment period. The effect size is −0.324 (−0.497) for the coverage loss measured on I/B/E/S (analyst reports). Thus, the measured effect amounts to 0.324 (0.497) standard deviations of the dependent variable.

Our results on coverage show that the quantity effect applies not only to EPS forecasts and recommendations from I/B/E/S but also to actual analyst reports. Replicating previous research on coverage proves that our sample of companies, dataset, and research design are suitable for investigating the effects of MiFID II.

6.2 Optimism

The subsequent analyses are based on the broker–company balanced sample, as explained in the methodology section. This reduces our sample to 55,184 analyst reports.

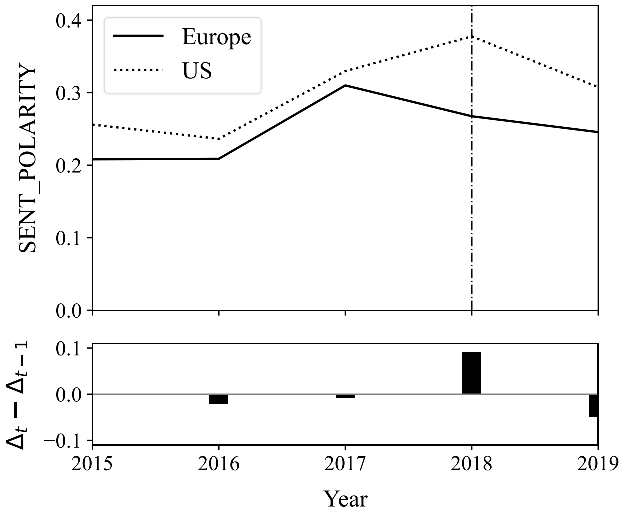

Figure 23. Changes in Sentiment Polarity Over Time

Figure 23 illustrates that with the introduction of MiFID II, the polarity in the sentiment of analyst reports regarding European companies is reduced compared to that of US companies. However, this does not necessarily mean that the optimism bias is reduced, as it could also be attributed to a difference in companies' performance between the European and US sample. The analyst could account for this by writing a less or more positive analyst report. To prevent such a distortion, we consider the stock market performance in our regression analysis (see Table 13). Model [4], which controls for the stock market return of the past 60 trading days, shows that for European companies the SENT_POLARITY and thus optimism have decreased significantly by −0.078 (see POST_TREAT). Since this analysis is on the level of the individual reports, not only correlation across multiple reports about the same company but also correlation across multiple reports from the same broker might be a concern. To account for this, we apply multiway clustering (Cameron and Miller 2015) for the calculation of standard errors of the regression model [4]. The effect size based on Cohen's *d* amounts to −0.185.

Variables	[3] SENT_POLARITY	[4] SENT_POLARITY
POST_TREAT	−0.056**	−0.078*
	(−7.63)	(−2.57)
POST	0.068**	
	(12.45)	
TREAT	−0.030**	
	(−6.35)	
ACTIVE_FUNDS		0.122
		(0.49)
MTB		0.0003*
		(2.14)
DEBT_RATIO		−0.132
		(−1.28)
SIZE		0.002
		(0.03)
ROA		0.068
		(0.20)
PPE_RATIO		−0.211
		(−0.90)
LOSS		−0.119*
		(−2.60)
GDP_GROWTH		0.011
		(0.03)
ln(GDP_CAPITA)		0.103
		(0.23)
RETURN		0.671**
		(10.61)
VOLATILITY		−0.145*
		(−2.06)
INTERCEPT	0.275**	−0.813
	(80.34)	(−0.16)
Observations	55,184	55,184
Fixed Effects	No	Firm, Broker & Year
Clustering	No	Firm & Broker
Adj. R^2	0.007	0.162

** *p*-value < 0.01; * *p*-value < 0.05; *t*-values reported in parentheses

Table 13. Regression Results – Changes in Analyst Optimism

6.3 Novelty

Figure 24 depicts the NOVELTY_SCORE for US and European companies over time. The reports dealing with US companies show an overall higher NOVELTY_SCORE. Furthermore, the novelty score for both groups rises after the introduction of MiFID II. However, this effect is stronger for European companies than for US companies, for which the novelty score rises only slightly.

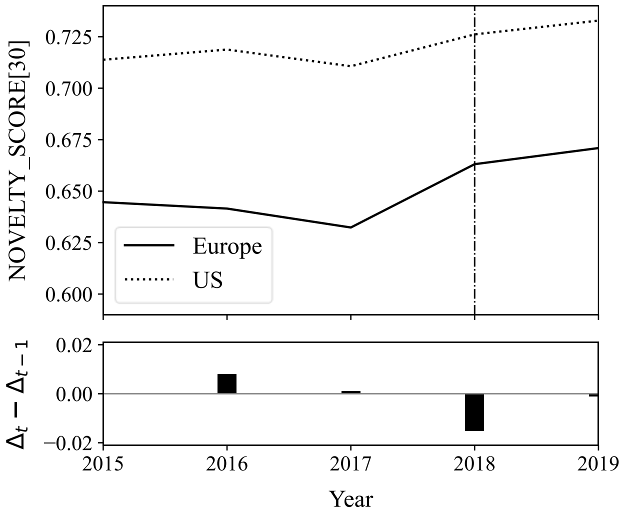

Figure 24. Changes in Novelty of Analyst Reports Over Time

The regression results in Table 14 confirm the insights gained from Figure 24. As in the previous analysis, the main focus is on the variable POST_TREAT. We apply three different model specifications. In the simplest model [5], without consideration of control variables and fixed effects, our variable of interest (POST_TREAT) shows a significant positive coefficient, which indicates that after the introduction of MiFID II, reports on European companies differ more strongly from previously published reports compared to the pre-MiFID II time frame. Additionally, the inclusion of fixed effects at the firm, broker, and year level and the consideration of key control variables [6] confirm this result. Even if the reference period is extended to 30 days [7], this result remains stable. Standard errors are clustered on the company and broker level from Model [6] onwards. These findings clearly demonstrate that the introduction of MiFID II has led analysts to provide more information not previously published by competing analysts to justify the price they must charge. We observe a Cohen's *d* of 0.130 for the 10-day NOVELTY_SCORE and 0.163 for the 30-day NOVELTY_SCORE. The measured effect size is slightly below the effect size from the analysis on optimism.

Variables	[5] NOVELTY_ SCORE[10]	[6] NOVELTY_ SCORE[10]	[7] NOVELTY_ SCORE[30]
POST_TREAT	0.009**	0.020*	0.023**
	(3.26)	(2.34)	(2.99)
POST	0.017**		
	(7.79)		
TREAT	−0.070**		
	(−37.91)		
ACTIVE_FUNDS		0.028	0.023
		(0.43)	(0.37)
MTB		−0.00001	0.000002
		(−0.28)	(0.07)
DEBT_RATIO		−0.015	−0.002
		(−0.54)	(−0.07)
SIZE		−0.003	−0.005
		(−0.16)	(−0.31)
ROA		0.091	0.097
		(1.25)	(1.44)
PPE_RATIO		0.078	0.111
		(1.07)	(1.53)
LOSS		0.002	0.005
		(0.17)	(0.61)
GDP_GROWTH		0.224	0.421*
		(1.09)	(2.39)
ln(GDP_CAPITA)		−0.174	−0.130
		(−1.32)	(−1.02)
RETURN		−0.003	−0.008
		(−0.22)	(−0.78)
VOLATILITY		0.003	−0.017
		(0.13)	(−0.89)
INTERCEPT	0.700**	2.580	2.164
	(530.23)	(1.70)	(1.47)
Observations	46,964	46,964	52,117
Fixed Effects	No	Firm, Broker & Year	Firm, Broker & Year
Clustering	No	Firm & Broker	Firm & Broker
Adj. R^2	0.047	0.394	0.478

** p-value < 0.01; * p-value < 0.05; t-values reported in parentheses

Table 14. Regression Results – Novelty of Textual Analyst Reports

The analysis on herding behavior around important disclosure events (i.e., EAs) by utilizing COS[EARN_DATE] confirms our findings on the NOVELTY_SCORE. The restriction to EAs with at least three analyst reports leaves a sample of 1,636 EAs with a total of 20,763 analyst reports being published during these EA time frames. Figure 25 shows that the COS[EARN_DATE] for European companies increases after the introduction of MiFID II. For the control group, however, it remains relatively constant over time.

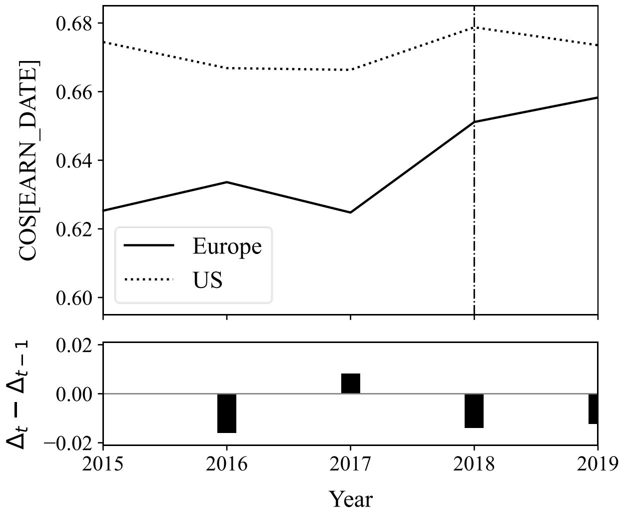

Figure 25. Changes in the Dissimilarity of Analyst Reports Issued During the Three Days Surrounding the Earnings Announcement

For the interaction term POST_TREAT, we observe a significant positive regression coefficient of 0.018 (see Table 15). Thus, analysts are more differentiating in terms of content when they publish reports compared to the pre-MiFID II period. Since the single observation is aggregated across multiple brokers, we only use firm and year fixed effects and cluster for robust standard errors on the firm dimension. The Cohen's *d* measures 0.169 for this analysis, which is about the same as the effect size observed on the NOVELTY_SCORE[30].

Variables	[8] COS [EARN_DATE]
POST_TREAT	0.018*
	(2.34)
ACTIVE_FUNDS	0.117*
	(2.00)
MTB	−0.0001**
	(−4.57)
DEBT_RATIO	0.008
	(0.18)
SIZE	−0.002
	(−0.15)
ROA	0.0318
	(0.37)
PPE_RATIO	−0.018
	(−0.26)
LOSS	0.007
	(0.62)
GDP_GROWTH	0.049
	(0.56)
ln(GDP_CAPITA)	−0.400**
	(−2.85)
RETURN	−0.019
	(−1.33)
VOLATILITY	0.007
	(0.31)
INTERCEPT	4.939**
	(3.07)
Observations	1,636
Fixed Effects	Firm & Year
Clustering	Firm
Adj. R^2	0.750

** *p*-value < 0.01; * *p*-value < 0.05
t-values reported in parentheses

Table 15. Regression Results – Dissimilarity of Analyst Reports Issued During the Three Days Surrounding the Earnings Announcement

6.4 Robustness Checks

The central assumption of the difference-in-differences approach is the parallel trend assumption (Angrist and Pischke 2008). It means that the control and treatment groups move in parallel except for the treatment effect. Thus, the control group should be a perfect counterfactual to the treatment group, and both groups are affected by external influences in the same way, except for the treatment. We use a placebo test to check whether the parallel trend assumption holds (Lechner 2011). For this purpose, we consider the pre-treatment period only (2015 until 2017). We then implement placebo treatments on January 1, 2016 and January 1, 2017. We perform the regression analysis for the coverage as shown in Table 12 (Model [1] and Model [2]), for the sentiment polarity in Table 13 (Model [4]), and the novelty scores as shown in Table 14 (Model [6] and Model [7]). The results for the placebo test are presented in Table 16. Only the 2017 placebo treatment on coverage shows a significant regression coefficient, which calls into question the results' causality that is based on this construct. However, it should be noted that our results on coverage are in line with previous research that has utilized larger samples (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021; Z. Liu and Yezegel 2020). Thus, the positive findings for the placebo treatment only limit the results on coverage to a limited extent. For all other models and placebo treatments, the test is negative (high *p*-values). This is a strong indication that the decrease in analyst optimism and the increase in novelty that we observe, and which are the main contributions of our paper, are driven by the introduction of MiFID II and, therefore, are causal. If the observations were due to a long-term trend and therefore not caused by MiFID II, significant results should also be expected for the placebo treatments (Guo and Mota 2021).

	[1]		[2]		[3]		[6]		[7]	
Dependent variables	COVERAGE [IBES]		[REPORTS]		SENT POLARITY		NOVETY_SCORE [10]		[30]	
Placebo Treatment	2016	2017	2016	2017	2016	2017	2016	2017	2016	2017
POST_TREAT	−1.390	−0.286	−1.170	1.750*	0.020	0.007	−0.021	−0.016	−0.012	−0.009
p-value	0.058	0.669	0.074	0.014	0.658	0.891	0.053	0.203	0.283	0.471
** *p*-value < 0.01; * *p*-value < 0.05										

Table 16. Placebo Testing

Due to its high accuracy, we use the FinBERT classifier to determine analyst optimism. As a robustness check, we also determine optimism based on the Loughran and McDonald (2011) word lists. This does not lead to changes in our results. As mentioned in the dataset section, our sample consists of approximately 14% of reports published by analyst firms that do not offer broker services themselves. If a company does not offer execution services and only sells research to the buy side, unbundling is unnecessary. As an additional robustness check, we conduct our analysis also solely with those analyst reports published by analyst firms that also offer brokerage services. The results remain unchanged.[1]

7 Discussion

Our study provides comprehensive insights into financial analysts' reactions to the changed market conditions caused by MiFID II. We focus on changes in the textual content of analyst reports. Both a quantity effect (coverage) and a quality effect (analyst optimism and novelty) can be observed in analyst reports. The number of analysts who publish an analyst report on European companies has decreased compared to that on US companies. At the same time, it can be seen that the analysts reduce their

[1] The results of these robustness checks are not tabulated but available upon request.

optimistic bias and differentiate themselves more strongly from the content of previously published reports. They also offer a more diverse information base around EAs. This finding is in line with previous research that has found increased stock market reactions following earnings forecasts after the implementation of MiFID II, as only new information should trigger stock market reactions (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021; Z. Liu and Yezegel 2020). Two possible channels of action could lead to more distinguishable analyst reports. The increased pressure due to the new pricing might have led to the fact that the analysts are more closely evaluated by the asset managers and therefore differentiate their reports more strongly from those of their competitors. However, this also requires that asset managers receive analyst reports from different brokers to carry out this evaluation. This argumentation is in line with the findings of Preece (2017), who found in his survey that investment professionals believe that the marketplace for investment research has become more competitive due to MiFID II. Another possibility is that the analysts simply have more time and resources to keep their reports up to date and differentiated due to the decreased quantity and increased focus. However, the setting used here does not allow us to evaluate which channel of action leads to the result.

Our analyses come with multiple limitations. We consider a relatively large sample in terms of the number of analyst reports (60,626 reports) but a relatively small one in terms of the number of companies (88 companies). However, this sample allows us to compare the contents of different reports, as analyst coverage is strong for these large companies.

Another limitation relates to the composition of the treatment and control groups. Since the effects of MiFID II expand beyond European borders (Allen 2019), it is not unlikely that, to a small extent, the treatment will also affect the control group. However, this would lead to an underestimation and not an overestimation of the MiFID II effects measured in our analysis. Furthermore, our treatment group does not consist of companies from all EEA countries. Due to the selection of the EURO STOXX 50 as a sample, only companies from the eurozone are included. However, as this also excludes companies located in the UK, the EURO STOXX 50 sample rules out the possibility that Brexit directly influences our results. However, it should be noted that MiFID II was also applied in the UK during the entire observation period.

In addition, the utilized data source has limitations. The brokerage houses’ decisions to publish their analyst reports via Thomson ONE are not random. Amiram et al. (2018) show that the overall forecast quality is higher in the Thomson ONE database than in I/B/E/S. They argue that brokerage houses with lower forecast quality have less incentive to offer their reports to Thomson ONE because of reputational costs. Thus, it is possible that the effect we measured is not due to an actual change in research quality but only due to a change in the selection of reports published via Thomson ONE. However, Amiram et al. (2018) attribute this selection effect in particular to a fundamental decision at the brokerage house level (brokerage houses with poor forecasting quality tend to publish their analyses in I/B/E/S but not their full reports in Thomson ONE). However, since we use a balanced sample in our analysis that only includes broker–firm combinations that are present before and after the treatment, and we include broker fixed effects in the main analyses, it seems unlikely that the results are solely due to the endogenous selection of the distribution channel.

The magnitude of the measured effects in this study is not only statistically but also economically significant. With a Cohen’s *d* of 0.497, the effect on analyst coverage can be described as an effect of medium strength, according to J. Cohen (1988). Although the effects of MiFID II on optimism and novelty can only be described as weak effects according to the effect size interpretation of J. Cohen (1988), the setting must also be taken into account when interpreting effect sizes. The heuristic separation between treatment and control leads to an underestimation of estimated effects and thus to underestimated effect sizes.

We use a state-of-the-art machine learning approach to measure the optimism of the analysts. This classifier has been trained and evaluated based on a sample of analyst reports. However, since such an approach also represents a black box to a certain extent, we conduct the same analysis based on the word lists from Loughran and McDonald (2011). Both methods lead to identical results. For measuring novelty, we use a proven method from the field of novelty detection (Yi Zhang et al. 2002). However, more advanced document representations, such as topic modeling or document embeddings instead of TF-IDF, are also conceivable, but we refrain from this to avoid further increasing the complexity of our analysis.

We assess analyst optimism based on sentiment polarity. However, there is no calibration point for this sentiment polarity. Therefore, it is impossible to make a statement based on the sentiment polarity alone regarding whether the report is too optimistic; we can only estimate whether it is rather optimistic or rather pessimistic. Unlike price targets, where a comparison between the estimated and observed value is possible, sentiment polarities do not offer this possibility. That said, our analysis of analyst optimism is based on the literature, which has shown that analysts are subject to an optimistic bias (Barber et al. 2006; Mehran and Stulz 2007). It can therefore be assumed that the text also tends to be written too optimistically. However, we can only conclude that optimism has decreased due to MiFID II; our approach cannot determine whether analysts have previously written text that was too positive when viewed objectively.

MiFID II came into force in January 2018 but was already passed by the legislature in 2014. This allowed market participants to anticipate adjustments. The European coverage development in I/B/E/S, as shown in Figure 22, provides an indication of a pre-trend that would lead to an underestimation of the MiFID II effect in our setting.

Our results have numerous implications for practitioners, interest groups, regulators, and researchers. First, we demonstrate that explicit pricing can improve quality by reducing analysts' behavioral biases of excessive optimism and herding behavior. Even if quality improvement is not the primary goal of the MiFID II regulation, it is still a positive side effect. The increased novelty implicates a decrease in the well-known herding behavior among financial analysts (Trueman 1994). The increased diversity of information can also improve market efficiency. Thus, our research explains the increased stock market reaction following individual analyst forecasts and recommendations (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021; Z. Liu and Yezegel 2020).

The European regulator can use these findings to evaluate the intended and unintended effects of MiFID II. However, it might also be of interest to regulators outside Europe. Allen (2019) shows that MiFID II has a global impact. Furthermore, a legal contradiction exists between the European MiFID II regulations and Section 202(a)(11) of the US Investment Advisors Act of 1940. The explicit billing for investment research required by MiFID II is only permitted to registered investment advisers under the Investment Advisors Act of 1940. This conflict is temporarily resolved by a No-Action Letter from the US Securities and Exchange Commission (SEC 2017, 2019). As this temporary exemption is only valid until June 3, 2023, the regulatory authorities must find a long-term solution to this problem. In its position paper, the CFA Institute advocates the unbundling of execution and research services in the US, thereby strengthening investors' transparency (Allen and Gellasch 2019). Our study can help regulators and interest groups in the US and worldwide to review and, if necessary, adapt their regulations and thus benefit from the experience gained within the EEA.

Furthermore, our results can help asset managers decide whether to continue purchasing analyst reports from the sell side or to produce research in-house. The motive for in-house research is to produce exclusive information. At the same time, our analyses suggest that sell-side analysts also try to provide more exclusive information by differentiating themselves more strongly from competitors. However, if the analyst report is sent to many buy-side companies, the informational value will decrease quickly, even if the reports contain a large portion of novel information.

Our research may provide a starting point for further research questions. The texts of analyst reports provide many additional opportunities to investigate the effects of MiFID II. Starting from the basic assumption that the coverage loss has led to an increased focus and more resources for creating individual analyst reports, it should be examined whether the weighting between information discovery and interpretation (A.H. Huang et al. 2018) has shifted due to MiFID II.

8 Conclusion

The regulatory change brought about by MiFID II has led to a major change in the marketplace of investment research. While investment research was previously cross-subsidized by the revenues from execution services, MiFID II requires the unbundling of execution and research services and has thus led to explicit pricing for analyst reports. In the run-up to the introduction of MiFID II, concerns were expressed that explicit pricing would reduce the demand for investment research and put sell-side analysts under pressure. Current research confirms these concerns and finds that analysts have withdrawn from less lucrative areas and that analyst coverage has decreased. However, this negative effect was also accompanied by a positive effect of increased quality of EPS forecasts. Our results show a decrease not only in the quantity of EPS forecasts but also in that of textual analyst reports. The main finding of our study is the reduced analyst optimism and herding behavior after the implementation of MiFID II. Both lead to more informative analyst reports. This finding is of interest both for European regulators and for regulators worldwide who consider reforming the regulation of investment research. Due to the improved quality of analyst reports, buy-side practitioners should consider continuing to buy sell-side analyst reports despite explicit pricing. The reports are written less optimistically and therefore provide a more objective picture. Most importantly, analyst reports contain more new and private information and carry a higher informational value than before MiFID II that the buy side could convert into successful asset management strategies. With the increased diversity of analyst reports, it tends to be more reasonable to purchase and read the reports of several analysts than before MiFID II.

Appendix A – Variable Definition

Variable	Definition	Source
COVERAGE [IBES]	Number of analysts forecasting the firm's earnings per share	Eikon, I/B/E/S
COVERAGE [REPORTS]	Number of brokers covering the firm by publishing an analyst report	Thomson ONE
SENT_POLARITY	The polarity between positive and negative sentences according to the FinBERT sentiment classifier pre-trained and finetuned by Y. Yang et al. (2020)	Thomson ONE, own calculation
NOVELTY_SCORE [10]	Mean cosine distance between the respective analyst report and all analyst reports issued in the last 10 days by other brokers than the respective analyst report	Thomson ONE, own calculation
NOVELTY_SCORE [30]	Mean cosine distance between the respective analyst report and all analyst reports issued in the last 30 days by other brokers than the respective analyst report	Thomson ONE, own calculation
COS[EARN_DATE]	Mean cosine distance between all reports published during the three days surrounding the conference call	Thomson ONE
POST	Binary variable set to 1 for all observations that occurred after the implementation of MiFID II (January 3, 2018) and 0 for the remaining observations	Thomson ONE, Eikon, I/B/E/S
TREAT	Binary variable set to 1 if the company has been a constituent of the EURO STOXX 50 during the observation period and 0 if it has been a constituent of the Dow Jones Industrial Average	Datastream
ACITVE_FUNDS	Number of shares held by actively managed funds divided by the number of shares held by all funds	Eikon
MTB	Market value of common equity divided by book value of common equity (Worldscope item 03501)	Datastream
DEBT_RATIO	Worldscope item 08221	Datastream
SIZE	Natural logarithm of the firm's total assets (Worldscope item 02999) converted to USD	Datastream
LOSS	Takes the value of 1 if the net income before extraordinary items or preferred dividends (Worldscope item 01551) is negative and 0 if positive	Datastream
ROA	Worldscope item 08326	Datastream
PPE_RATIO	Book value of property, plant, and equipment (Worldscope item 02501) divided by total assets (Worldscope item 02999)	Datastream
RETURN	Stock return during the last 60 trading days prior to the observation	Datastream
VOLATILITY	Annualized volatility of the daily stock returns during the last 60 trading days prior to the observation	Datastream

Table 17. Variable Definition

Appendix B – Preprocessing of Analyst Reports

The analyst reports are provided as PDF files. We extract the text from these files by transforming them into a tabular structure. Later we remove any graphs, tables, and boilerplate (e.g., disclaimers) (A.H. Huang et al. 2018). This procedure allows us to extract the analyst report's textual essence for the subsequent analysis. We apply a sentence tokenizer to separate the sentences and transform characters to lower case. The sentences are used for the sentiment analysis (SENT_POLARITY). For subsequent analyses, we perform further standard pre-processing steps. All non-ASCII characters are removed, and the text is tokenized into single words. The tokens are then reduced to their word stem utilizing the Porter stemmer (Porter 1980). We drop stop words, which are frequently occurring words that typically do not add information to the text (e.g., articles), as well as short words consisting of less than three characters. Finally, we use a TDM for text representation. In this text representation, each unique word within the corpus is represented by a single column, and each analyst report is represented by a single row. We reduce the feature set (number of columns) of the TDM by removing words that occur in less than 0.5% or more than 30% of all analyst reports. This feature reduction preserves the words with high information value for our analysis (Farrell et al. 2018). The vectors of the TDM are used to calculate the NOVELTY_SCORE.

II.3. MiFID II and Media Coverage

MiFID II and its Unintended Effects on Media Coverage

Authors: Steffen Bankamp

Outlet: Working paper, prior version accepted and presented at the 12th Annual Pre-ICIS Workshop on Accounting Information Systems, Austin, United States, 2021 and at the 44th Annual Congress of the European Accounting Association, Bergen, Norway, 2022.

Abstract: The introduction of a new European regulation called MiFID II had a substantial impact on capital market participants. Building on principal–agent theory, this paper examines the interrelation of two important information intermediaries in the capital market (sell-side analysts and journalists) who are asymmetrically regulated by MiFID II. While the effects on sell-side analysts targeted directly by MiFID II have been extensively investigated by prior research, the existence of indirect effects on non-regulated information intermediaries remains unclear. This paper provides evidence that the research unbundling required by MiFID II also negatively affects media coverage. This applies only to full articles, not to flash articles. It is also apparent that media coverage is particularly declining during the period of earnings announcements. Furthermore, the results show that this effect originates from the channel of reduced analyst coverage, making it more expensive for journalists to create information.

Keywords: MiFID II, regulation, media coverage, analyst coverage, research unbundling

1 Introduction

On capital markets, many principal–agent relationships are observable, such as the one between a company's management (agent) and its shareholders (principal) (Jensen and Meckling 1976). This is typically characterized by information asymmetries in favor of the management (Richardson 2000). Information asymmetries leads to agency costs that can be reduced by mitigating these asymmetries. Information intermediaries such as financial analysts or media can help to reduce these information asymmetries in the capital market (Drake et al. 2014; Mola et al. 2013).

Financial media makes an essential contribution by disseminating information to the capital market (L. Fang and Peress 2009; Bushee et al. 2010; Drake et al. 2014). Media also reaches small and private investors who do not have access to exclusive information (e.g., access to analyst reports or direct access to the management team). At the same time, financial journalists rely on sell-side analysts as an important source for their articles (Li 2015; Call et al. 2022; Rees et al. 2015). The introduction of Directive 2014/65/EU (EU 2014), known as "Markets in Financial Instruments Directive II" (MiFID II), in Europe on January 3, 2018 had a major impact on the capital market and on financial analysts in particular. The latter are no longer allowed to follow the industry practice of selling investment analyses as a bundle with execution services. Instead, they are required to price investment research explicitly (unbundling), which has led to changes in entire business models and consequently to a decline in analyst coverage for European companies (e.g., B. Fang et al. 2020). Studies on liquidity in the capital market also indicate that MiFID II has worsened the overall information environment (Lang et al. 2021; B. Fang et al. 2020; Amzallag et al. 2021). While the existing literature has extensively researched the impact on analysts' information production (B. Fang et al. 2020; Lang et al. 2021; Z. Liu and Yezegel 2020; Guo and Mota 2021; Anselmi and Petrella 2021; Amzallag et al. 2021; Preece 2017, 2019), the potential impact on financial journalists has remained unexplored. Although, unlike analysts, journalists are not directly affected by MiFID II, they could still be indirectly affected by the reduced analyst coverage. Guest and Kim (2022) find that a decline in analyst coverage negatively influences media coverage. Against the background of MiFID II's objective to increase transparency in the capital market and protect investors, it is crucial to examine the unintended effects of the regulation that could counteract its objectives by limiting media's ability to reduce information asymmetry. This leads to the overall research question (**RQ)** of this paper.

RQ: *How does MiFID II's research unbundling impact the provision of information by the media?*

Despite fundamental research by Guest and Kim (2022) that demonstrates a positive impact of analyst coverage on media coverage, it is important to examine this relation against the background of MiFID II, as this setting comes with different underlying conditions. First, the study by Guest and Kim (2022) is based on the setting of brokerage mergers. This means that although the merger reduces the number of reports, the quantity of information might not be affected, as private information from both brokerage houses may still be incorporated in the new report of the merged broker. This is different in the MiFID II setting of this study: the regulation creates a situation in which brokerage houses explicitly stop covering individual companies, but this means that the private information of the analyst who stops covering cannot be transmitted via analyst reports of peers. In addition, brokerage mergers are driven by synergy considerations on the broker level, whereas the coverage decision following the regulatory change is driven by considerations of added value and focus (Lang et al. 2021). In addition, the study by Guest and Kim (2022) is based on a dataset ranging from 2000 to 2009 and thus mainly from the period before the financial crisis. The financial crisis and the responses to it, including MiFID II itself, have led to many changes in the capital market. For all these reasons, it is necessary to analyze the impact of MiFID II on sell-side analysts, financial journalists, and their interactions by using contemporary data. Effects on the information provision of financial journalists are of particular relevance, as they are an important information source for individual investors. In a recent literature review, Bender et al. (2021) call for more research on the bundling and unbundling of investment research and execution services.

The results presented in this paper show that MiFID II negatively impacts both the quantity of analyst output and the quantity of media output. In this context, the analysis indicates that this relationship is causal and that MiFID II affects journalists via the transmission channel of reduced analyst coverage. In addition, media coverage during earnings announcement (EA) time frames is more strongly affected from MiFID II compared with media coverage in other time periods. The observed reduction is statistically and economically significant, as media coverage drops by 12% to 7%, depending on the utilized sample. Media coverage declines only for full articles (articles with a body text) but not for flash articles (headlines only). Concerning company size, no heterogeneous impact of MiFID II on media coverage can be identified. Overall, the results indicate that after the introduction of MiFID II, journalists contribute significantly less to the reduction of information asymmetries on the capital market.

This paper contributes to the research strand that examines the effects of the research unbundling required by MiFID II (e.g., B. Fang et al. 2020). In doing so, the paper contributes by examining the directly affected information intermediaries (sell-side analysts) and observing indirect effects on other information intermediaries (journalists). This study also contributes to the research strand dealing with the connection between media and sell-side analysts (e.g., Call et al. 2022). Using a contemporary setting, the results of Guest and Kim (2022) can be confirmed. Reduced analyst coverage leads to reduced media coverage.

This study is of high practical relevance, as the results can be considered when revising the current MiFID II regulation. Recently (February 2021), the European Parliament loosened the MiFID II regulation and now allows the bundling of investment research about small and medium-sized companies.

2 Institutional Background and Hypothesis Development

2.1 Institutional Background on MiFID II

The European Parliament passed the new MiFID II regulation in 2014, which market participants have had to apply since January 3, 2018. MiFID II is the successor of the MiFID regulation (Directive 2004/39/EC; EU (2004)), which had been in force since 2007. MiFID II responds to weaknesses in the financial system that became apparent during the financial crisis. The harmonized rules are intended to reduce barriers for market participants to provide and demand financial services from all over Europe, thereby creating a single European financial market. The regulation focuses in particular on increasing transparency and protecting investors. To achieve these goals, the new regulation includes a wide range of rules on, inter alia, multilateral trading facilities, algorithmic trading, and corporate governance. In addition, research unbundling, which is extensively discussed in this paper, is also part of MiFID II.

The MiFID II regulation is applicable for a financial service provider (e.g., investment firm or market operator) that offers financial services or conducts investment activities through a branch in the European Economic Area (EEA) (Directive 2014/65/EU Article 1(1); EU (2014)). Thus, MiFID II is mandatory for all investment companies based in EEA's member states.[2] However, there is a particular aspect concerning the applicability of research unbundling: the regulation does not directly target the brokerage houses that produce investment research but rather the purchaser of these services. Thus, asset managers who receive both trading and research services from a broker must insist that the broker separately charges for trading and research services. This also means that brokers who are not domiciled within the EEA but have European asset managers as clients may be indirectly affected by MiFID II.

[2] EEA member states are all countries of the European Union, Iceland, Liechtenstein and Norway. The United Kingdom had to apply EEA regulations until the end of the Brexit's transition period (2020-12-31).

In response to the COVID-19 pandemic, a package of measures was passed in February 2021 that provides numerous easements to the MiFID II regulations (Directive (EU) 2021/338; EU (2021)). These include easements on research unbundling. Asset managers are again permitted to source investment research on small and medium-sized companies with a market capitalization of up to €1 billion in a bundle with execution services.

2.2 MiFID II and its Effect on Analyst Research Output

The question of how the bundling of execution and research services affects investors has long preoccupied regulators, practitioners, and researchers alike. Bender et al. (2021) provide a comprehensive literature review on papers that have addressed this issue. With the introduction of MiFID II, the European Parliament has stopped the practice of bundling. Since this regulation fundamentally changes the business model of analysts, the new regulation had already attracted much attention even before it came into effect. The CFA Institute conducted a study in advance to assess the potential impact of the research unbundling demanded by the European Parliament (Preece 2017). The investment professionals who were surveyed provided a clear picture of their opinion. The majority (78%) of buy-side professionals expect to reduce the amount of research procured from the sell side (Preece 2017). At the same time, it is assumed that in-house analysts will provide more research. With this considerable decline in demand on the buy side, the question arises as to whether the sell side can maintain its information supply and to what extent this might affect analyst coverage.

Shortly after MiFID II came into force, the first quantitative studies on the impact of MiFID II on analysts' research output were published (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021; Z. Liu and Yezegel 2020; Amzallag et al. 2021; Anselmi and Petrella 2021). The focus of these studies is on the impact on the quantity and quality of analysts' research. All six studies confirm a coverage loss due to MiFID II. This effect can therefore be regarded as relatively certain. However, the authors come to different conclusions regarding company characteristics associated with the magnitude of the MiFID II-induced coverage loss: Guo and Mota (2021), Lang et al. (2021), Amzallag et al. (2021), and Anselmi and Petrella (2021) conclude that large companies are particularly affected by the coverage loss, whereas B. Fang et al. (2020) find that small companies are more affected than large companies.

In terms of the quality of analyst output, most studies find a positive effect of MiFID II (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021; Z. Liu and Yezegel 2020). This is particularly reflected in lower earnings-per-share (EPS) forecast errors. Only the study of Amzallag et al. (2021) could not confirm a change in analysts' research quality.

Given that MiFID II has reduced the quantity of sell-side research while increasing its quality, the question arises as to how this has affected the overall information environment within the capital market. In line with the increased quality, researchers have observed a stronger market reaction to the individual analyst forecast (B. Fang et al. 2020; Guo and Mota 2021; Lang et al. 2021; Z. Liu and Yezegel 2020). On an aggregated level, however, the forecasts appear to be less informative due to the decrease in quantity. A large proportion of the studies have found a decrease in liquidity, which indicates higher information asymmetries (Amzallag et al. 2021; B. Fang et al. 2020; Lang et al. 2021). The studies have thus concluded that the negative quantity aspect overrides the positive quality aspect. However, Guo and Mota (2021) and Anselmi and Petrella (2021) have come to a different conclusion, as they did not observe any effects of MiFID II on bid-ask spreads – except a negligible effect on micro-stocks in Anselmi and Petrella (2021) – and conclude that MiFID II did not affect market liquidity. Anselmi and Petrella (2021) even argue that MiFID II merely led to a reduction in the overprovision of analyst output in Europe.

2.3 Relation Between Analyst Coverage and News Coverage

Prior research has called for an investigation of the interrelation between media and financial analysts (G.S. Miller and Skinner 2015). The fundamental question is whether the output of analysts and journalists is complementary or a substitute. If the two products are complementary (substitute), a positive (negative) relationship is expected. L. Fang and Peress (2009) found that the media coverage of stocks covered by analysts is lower and thus argued that media and analyst outputs are substitutes. Bushee et al. (2010) and Drake et al. (2014) used the number of analysts as an explanatory variable in a multivariate regression to explain media coverage. In both studies, analyst coverage was found to positively impact media coverage, which speaks in favor of them being complementary goods. Guest and Kim (2022) are the first to explore the relation of media coverage and analyst coverage in a way that allows them to make statements about the causality of this relationship by exploiting a natural experiment. They use a setting where analyst coverage decreased due to the mergers of brokerage houses. Both brokerage houses covered the respective company before the merger, while only the newly built brokerage house covers the company after the merger. According to the authors' findings, this decrease also leads to a decrease in media coverage, as fewer articles have been published. The authors find that this effect is especially strong on companies that have overall lower analyst coverage, are not listed in the S&P 500, have a more complex disclosure, and rated the lost analyst as an all-star analyst. The authors argue that the cost of writing articles increases with less information available from financial analysts, making it less worthwhile to write articles. This effect, which the authors also describe as a cost-dominated hypothesis, outweighs a potentially positive effect resulting from the reduced competition with financial analysts for readers. Call et al. (2022), who surveyed 462 financial journalists, also found that financial journalists extensively use analysts as an information source. This is in line with Kuperman et al. (2003) and Li (2015) findings that a substantial proportion of news articles in the Wall Street Journal reference analysts in their articles.

Overall, the evidence from previous research suggests that analysts are important information providers for the media. Thus, a decrease in analyst coverage may negatively impact the information base of those who consume analyst reports directly and those who indirectly consume this information through the vehicle of news media. This leads to the main hypothesis (**H1**):

H1: *MiFID II's research unbundling leads to the unintended effect of reduced media coverage.*

As highlighted in the literature section, the influence of firm size on the MiFID II effects is still unclear, and researchers have differing conclusions in this regard (e.g., Amzallag et al. 2021; B. Fang et al. 2020). Concerning media coverage, this study assumes that journalists will reduce coverage of small companies less, as the majority of prior research indicates that small companies are also less affected by analyst coverage losses from MiFID II. Following the argumentation of reduced media coverage being transmitted via the channel of analyst coverage, the media coverage loss for small companies should also be weaker. This leads to the second hypothesis (**H2**):

H2: *Small companies are less affected by the media coverage loss than large companies.*

Guest and Kim (2022) split media coverage into full articles and flash articles, which consist solely of a headline. Based on their observations, it will be assumed that the loss is particularly high in the case of full articles, since journalists must research content for these articles, which is more difficult without input from analysts. In the case of flash articles, journalists need significantly less background information. This leads to the third hypothesis (**H3**):

H3: *MiFID II's research unbundling leads to stronger coverage loss of full than of flash articles.*

Finally, it should be clarified whether MiFID II has led to a shift in the timing of media coverage. Guest and Kim (2022) argue that journalists are increasingly focusing on EA dates when analyst coverage is reduced and producing articles around these dates, as it is easier to write articles without analysts' input if they can use the information revealed by the company on the announcement date. At the same time, A.H. Huang et al. (2018) show that analyst reports cluster strongly around EA dates. In contrast to the argumentation of Guest and Kim (2022), this would mean that a decline in analyst coverage would especially impact this period. As a result, a reduction in media coverage should be more pronounced during this main period of analyst output, which builds the fourth hypothesis (**H4**):

H4: *The loss in media coverage is especially pronounced during the time of EAs.*

3 Research Design

To answer the research question of how MiFID II has affected media coverage and, in particular, the information environment of retail investors, a three-stage research design is conducted.

As a first step, the sample is validated by testing the well-established result of reduced analyst coverage. For this purpose, a difference-in-differences approach is applied, as done in previous studies on the effects of MiFID II (B. Fang et al. 2020; Lang et al. 2021; Guo and Mota 2021; Amzallag et al. 2021; Z. Liu and Yezegel 2020; Anselmi and Petrella 2021).[3] This allows assessing whether and how the dependent variable (analyst coverage) in the treatment group (EEA) has changed compared to the control group (US) as a result of the treatment (introduction of MiFID II). This allocation of the treatment and control group does not correspond precisely to the regulatory reference point of MiFID II, as the location of the asset manager who procures the research and execution services determines whether unbundling is mandatory. However, since it is not known which analyst reports are consumed by which asset managers, the company location offers a suitable alternative. Due to the home bias in investment decisions (e.g., Coval and Moskowitz 1999), it is to be expected that European asset managers prefer to invest in European companies and accordingly demand research on those companies. Guo and Mota (2021) have already followed this heuristic and used the home bias to justify allocating the treatment and control groups at the company level. The difference-in-differences approach is practical when the treatment is not assigned on an individual (e.g., company) level but on a higher level and thus affects entire regions or groups as a whole (Angrist and Pischke 2008). This is particularly the case if changes in legislation, regulation, or other policies are observed that relate to a specific jurisdiction. This setting controls for all time-invariant omitted variables at the group level (EEA vs. US) as well as omitted variables that lead to trends affecting both groups simultaneously and equally. These include, for example, general developments

[3] The authors Anselmi and Petrella (2021) have conducted this difference-in-difference approach only between SME companies and large companies from Europe. All other studies have compared European companies with companies from outside of Europe (e.g. North America).

in the capital market, such as the trend toward passive investment products (Anadu et al. 2020), that could influence analyst output. The basic idea behind the difference-in-differences approach is that, without the introduction of MiFID II, developments within the EEA would have been parallel to those in the US. If an offset emerged in this parallel development at the time of the introduction of MiFID II, this effect is attributable to the introduction of MiFID II (Angrist and Pischke 2008).

In a second step, the impact of a MiFID II-induced analyst coverage loss on the media coverage of these companies is evaluated. Therefore, EEA companies that lost analyst coverage after the MiFID II implementation are compared with EEA companies experiencing a stable coverage. This structure allows the effect to be evaluated using a difference-in-differences approach.

While the preceding analysis indicates whether MiFID II impacted news coverage via analyst coverage, the problem of endogeneity still exists. Omitted variables could drive analyst and media coverage simultaneously. Even though control variables for numerous factors are applied, the possibility of other omitted variables driving analyst and media coverage cannot be ruled out.

Guest and Kim (2022) use a natural experiment (brokerage mergers) to overcome endogeneity when linking analyst coverage to media coverage. In the present study, the ordinary difference-in-differences approach with EEA companies as the treatment and US companies as the control group is applied to evaluate the effect of MiFID II on media coverage. This approach reduces the problem of endogeneity. Shifts, where individual companies have become more or less important to analysts and/or journalists, should balance one another out at the overall market level. This is a strength of the difference-in-differences setting comparing the entire EEA market to the US market which makes the analysis more robust.

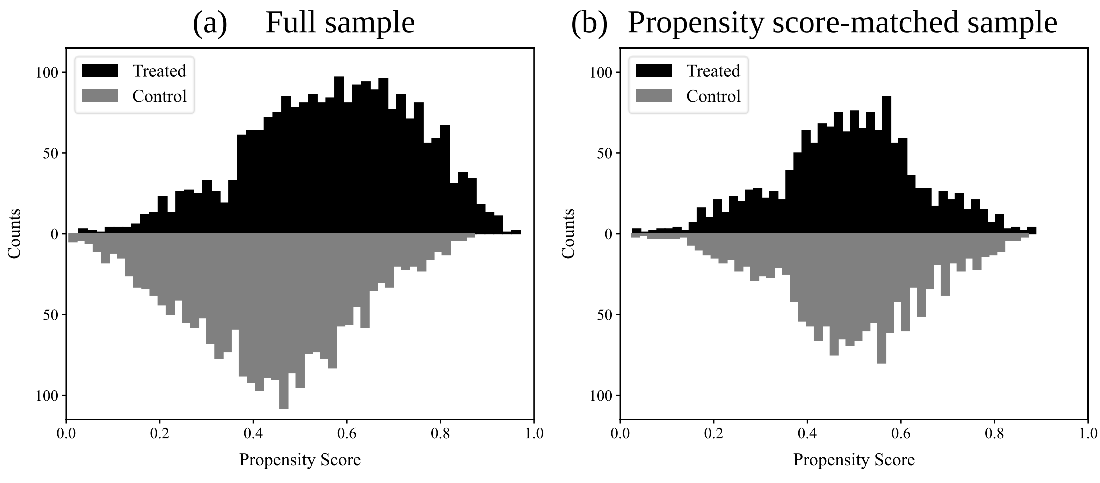

Figure 26. Distribution of the Propensity Score

To reduce the bias between the treatment (EEA sample) and the control group (US sample) in the pre-treatment period, propensity score matching (PSM) is applied (Rosenbaum and Rubin 1983). The control company from the same TRBC economic sector with the closest propensity score (without replacement) is matched to each company in the treatment group. A caliper of 0.184 is applied, corresponding to 0.2 standard deviations of the propensity score (Austin 2011). This defines the maximum accepted distance of the propensity scores at which matching is performed. The propensity score is calculated based on all control variables of the year 2016. The matching reduces the sample to 11,808 firm-year observations. Figure 26 illustrates that the propensity score matching reduces the difference in the distribution of the propensity scores between the two groups. For the pre-MiFID II era, all control variables except for PPE_RATIO show a significant difference (1% level) between the US and the EEA sample. By applying propensity score matching, only four of the 11 control variables used in the subsequent analysis are significantly different (1% level) between these two groups.[4]

3.1 MiFID II and Analyst Coverage

Equation (6) is estimated to calculate MiFID II's effect on the analyst coverage of European companies. Each observation is on the company-year level, where i is the index for the company and t the index for the respective year:

$$\begin{aligned}
&\mathrm{N_ANALYSTS}_{t,i} = \alpha_0 + \alpha_1 \mathrm{POST_TREAT}_{t,i} + \alpha_2 \ln(\mathrm{MARKETCAP}_{t,i}) \\
&+\alpha_3 \ln(\mathrm{TOTAL_ASSETS}_{t,i}) + \alpha_4 \mathrm{PPE_RATIO}_{t,i} + \alpha_5 \mathrm{DEBT_RATIO}_{t,i} \\
&+\alpha_6 \mathrm{MTB}_{t,i} + \alpha_7 \mathrm{LOSS}_{t,i} + \alpha_8 \mathrm{ROA}_{t,i} + \alpha_9 \mathrm{TOTAL_RETURN}_{t,i} \\
&+\alpha_{10} \mathrm{VOLATILITY}_{t,i} + \alpha_{11} \mathrm{TRADING_VOLUME}_{t,i} \\
&+\alpha_{12} \mathrm{EPS_DISPERSION}_{t,i} + \alpha_i + \alpha_t + \varepsilon_{t,i}
\end{aligned} \tag{6}$$

The analyst coverage (N_ANALYSTS) is measured as the number of sell-side analysts submitting an EPS forecast to I/B/E/S during the respective calendar year. The interaction term POST_TREAT is the main variable of interest, as its regression coefficient α_1 shows if MiFID II affects analyst coverage. The interaction term is set to 1 for all company-years of European companies after the introduction of MiFID II (2018 and 2019). For all company-years of US companies and all European company-years in the pre-MiFID II era, the interaction term remains 0.

Building on existing research, numerous control variables at the company-year level that could influence analyst coverage are applied (e.g., B. Fang et al. 2020; Mola et al. 2013). The natural logarithms of MARKETCAP and TOTAL_ASSETS are applied to

[4] T-test for the pre-treatment differences of the EEA and US sample are not tabulated for brevity but are available upon request.

account for the company size, both measured in million USD. Mola et al. (2013) show that total assets are negatively associated with analyst coverage loss. Moreover, PPE_RATIO is used as a control for the capital intensity of the underlying business (Aerts et al. 2008). The control variable DEBT_RATIO accounts for the difference in capital structure and can serve as a measure for financial distress, which has been found to be negatively associated with analyst coverage loss (Mola et al. 2013). To account for the relative firm's valuation, the variable MTB (market to book ratio) is used. For firm and stock market performance, the variables LOSS, ROA, and TOTAL_RETURN are used. These variables are important as poorly performing firms might hold back information and make it harder for analysts to conduct their analysis (S.Y. Ho et al. 2019; Kothari et al. 2009). Furthermore, VOLATILITY is used to account for the firm's risk. Lang et al. (2021) find that volatility is positively related to analyst coverage. TRADING_VOLUME, which is scaled by the company's total shares, is used to account for differences in trading activity. To account for the heterogeneity of analysts' beliefs on a specific stock, the variable EPS_DISPERSION is used (Diether et al. 2002). This variable is constructed as the standard deviation of all EPS forecasts of the respective firm-year divided by the median of the forecast. The average EPS_DISPERSION of the TRBC business sector is assigned to the firm-year if fewer than three EPS forecasts are available (Guest and Kim 2022). A detailed description of all variables, including the data sources and their construction, can be found in Table 26 (Appendix). In addition, the company fixed effects α_i control for all company-specific characteristics that are fixed over time (e.g., industry). The same applies to year-specific effects that simultaneously affect all companies, which are controlled for by year fixed effects α_t.

3.2 Analyst Coverage Loss and Media Coverage

To determine how the loss of analyst coverage due to MiFID affects media coverage, first only the European sample is considered, since only for this sample can it be assumed that a coverage loss is caused by MiFID II. As already mentioned, only companies with stable analyst coverage before the introduction of MiFID II were included. The sample is divided into companies that show or do not show a coverage loss during the post-MiFID II era (2018 and 2019) compared to the pre-MiFID II era (2016 and 2017). For companies with a coverage loss, the variable COVERAGE_LOSS is set to 1. For all other companies, the value of the variable is set to 0. Companies without any analyst coverage during the pre-MiFID II era, companies with a fluctuation of more than 10% in analyst coverage during the pre-MiFID II era, and companies that experienced an increase in analyst coverage of more than 20% after the introduction of MiFID II are excluded from this analysis. These filtering steps increase the likelihood that an observed effect of the variable of interest COVERAGE_LOSS can be attributed to the analyst coverage loss caused by MiFID II. Media coverage is calculated as the natural logarithm of one plus the number of full articles published during period i (Guest and Kim 2022). To understand the impact on media coverage, Equation (7) is estimated, where the control variables are equal to those in Equation (6):

$$\begin{aligned}&\ln(1+\mathrm{N\_FULL\_ARTICLES})_{t,i}=\alpha_0+\alpha_1\mathrm{COVERAGE\_LOSS}_{t,i}\\&+\alpha_2\ln(\mathrm{MARKETCAP}_{t,i})+\alpha_3\ln(\mathrm{TOTAL\_ASSETS}_{t,i})+\alpha_4\mathrm{PPE\_RATIO}_{t,i}\\&+\alpha_5\mathrm{DEBT\_RATIO}_{t,i}+\alpha_6\mathrm{MTB}_{t,i}+\alpha_7\mathrm{LOSS}_{t,i}+\alpha_8\mathrm{ROA}_{t,i}\\&+\alpha_9\mathrm{TOTAL\_RETURN}_{t,i}+\alpha_{10}\mathrm{VOLATILITY}_{t,i}+\alpha_{11}\mathrm{TRADING\_VOLUME}_{t,i}\\&+\alpha_{12}\mathrm{EPS\_DISPERSION}_{t,i}+\alpha_i+\alpha_t+\varepsilon_{t,i}\end{aligned}\tag{7}$$

The variable of interest is COVERAGE_LOSS. The regression coefficient α_1 is expected to be negative, since the loss of analyst coverage should negatively impact the journalists' output.

3.3 MiFID II and Media Coverage

To investigate whether MiFID II causally impacts the media coverage of European companies, a difference-in-differences approach is employed again by using the US sample as a control group. Regression Equation (8) corresponds to the structure of Equation (6) and contains the same independent variables, with the dependent variable corresponding to that of Equation (7):

$$\begin{aligned}&\ln(1+\mathrm{N\_FULL\_ARTICLES})_{t,i}=\alpha_0+\alpha_1\mathrm{POST\_TREAT}_{t,i}\\&+\alpha_2\ln(\mathrm{MARKETCAP}_{t,i})+\alpha_3\ln(\mathrm{TOTAL\_ASSETS}_{t,i})+\alpha_4\mathrm{PPE\_RATIO}_{t,i}\\&+\alpha_5\mathrm{DEBT\_RATIO}_{t,i}+\alpha_6\mathrm{MTB}_{t,i}+\alpha_7\mathrm{LOSS}_{t,i}+\alpha_8\mathrm{ROA}_{t,i}\\&+\alpha_9\mathrm{TOTAL\_RETURN}_{t,i}+\alpha_{10}\mathrm{VOLATILITY}_{t,i}+\alpha_{11}\mathrm{TRADING\_VOLUME}_{t,i}\\&+\alpha_{12}\mathrm{EPS\_DISPERSION}_{t,i}+\alpha_i+\alpha_t+\varepsilon_{t,i}\end{aligned}\tag{8}$$

Based on **H1**, the interaction term POST_TREAT is expected to be negative.

4 Sample Selection and Descriptive Statistics

The initial company sample is obtained from Refinitiv Workspace. It consists of all public listed companies headquartered within the European Economic Area[5] (United States) for the treatment (control) group. Similar to Guo and Mota (2021), for the subsequent analysis, only companies are considered for which control variables

[5] All member states that are members of the EEA during the entire observation period are taken into account. For this reason, the United Kingdom is also considered, as Brexit did take place in January 2020 after the observation period of this study.

(accounting and price data) are obtainable for the entire observation period ranging from 2016 until 2019. This prevents IPOs or delistings from affecting the results. Additionally, only companies are considered that occur at least once during the observation period in I/B/E/S and are mentioned in RavenPack's entity mapping file and can thus be detected by RavenPack. Detailed information on the sample construction, including the quantitative effects of the filters described above, are tabulated in Table 18.

Sampling Steps	*N* Firms			*N* Firm-Years		
	Overall	EEA	US	Overall	EEA	US
Firms from Workspace Screener	18,198	8,979	9,219	72,792	35,916	36,876
Less firms without any I/B/E/S EPS forecast during the observation period	-11,046	-5,296	-5,750	-44,184	-21,184	-23,000
Less firms not an entity in Raven Pack	-114	-60	-54	-456	-240	-216
Less firms without accounting and price data	-1,033	-614	-419	-6,953	-3,810	-3,143
Less firms without full history from 2016 to 2019	-1,426	-710	-716	-2,883	-1,486	-1,397
Full Sample	**4,579**	**2,299**	**2,280**	**18,316**	**9,196**	**9,120**
Less firms dropped in PSM	-1,627	-823	-804	-6,508	-3,292	-3,216
PSM Sample	**2,952**	**1,476**	**1,476**	**11,808**	**5,904**	**5,904**

Table 18. Sample Construction

To measure analyst coverage, the number of analysts that have provided an EPS forecast in I/B/E/S is used. The data on news coverage stems from RavenPack, a proprietary data provider that automatically analyses a large variety of news stories from numerous sources and transforms the unstructured data into a tabular structure (RavenPack 2021). RavenPack, therefore, detects entities (e.g., companies, countries, people, or products), events, and topics within these stories and calculates different figures (e.g., sentiment scores) that represent the news story or the event detected. All news stories about the entire company sample that have been posted during the observation period are downloaded from RavenPack. The media coverage is calculated as the number of stories that have been published about the specific company during the respective year. Following Guest and Kim (2022), this variable is calculated for full articles (N_FULL_ARTICLES) and flash articles (N_FLASH_ARTICLES) separately. Full articles are ordinary articles with a headline and at least one further paragraph, whereas flash articles only consist of a headline without any body text. To improve data quality and reduce noise, two filters are applied. First, only news stories are considered, where the related company has a maximum relevance score of 100. RavenPack assigns this score if the company is detected as the first entity in the headline. This filter ensures that news articles in which a company is only marginally mentioned are not counted as news coverage. Thus, only those articles in which the company in question plays the main role are considered. To prevent new news sources that have been added to

RavenPack over time from influencing the analysis, only those news articles that originate from sources already covered by RavenPack before the start of the observation period are considered. Balance sheet data and capital market data are obtained from Refinitiv Datastream and Refinitiv Workspace to construct the control variables in this study. All utilized variables and information on their construction can be found in Table 26 (Appendix). All continuous variables are winsorized at the 1% and 99% quantile to prevent outliers from distorting the results. The descriptive statistics on the variables after winsorizing are listed in Table 19.

Variable	***N***	**Mean**	***SD***	**P25**	**Median**	**P75**
N_ANALYSTS	18,316	8.180	7.865	2.000	5.000	12.000
N_FULL_ARTICLES	18,316	144.131	229.783	9.000	79.000	178.000
N_FLASH_ARTICLES	18,316	124.756	97.765	41.000	115.000	186.000
N_FULL_ARTICLES_EA	18,316	41.404	67.161	2.000	22.000	51.000
N_FULL_ARTICLES_NEA	18,316	102.923	171.704	6.000	52.000	126.000
N_FLASH_ARTICLES_EA	18,316	90.909	76.960	21.000	78.000	146.000
N_FULL_ARTICLES_NEA	18,316	33.739	37.930	8.000	22.000	46.000
MARKETCAP	18,316	5,898	15,395	213	925	3,791
TOTAL_ASSETS	18,316	11,360	37,164	235	1,244	5,484
PPE_RATIO	18,316	0.249	0.268	0.039	0.140	0.369
DEBT_RATIO	18,316	0.370	0.291	0.141	0.357	0.531
MTB	18,316	3.207	6.039	1.180	2.000	3.650
LOSS	18,316	0.252	0.434	0.000	0.000	1.000
ROA	18,316	-0.003	0.202	0.006	0.041	0.079
TOTAL_RETURN	18,316	0.113	0.438	-0.153	0.078	0.319
VOLATILITY	18,316	0.383	0.212	0.244	0.321	0.447
TRADING_VOLUME	18,316	0.548	0.509	0.230	0.426	0.685
EPS_DISPERSION	18,316	0.378	0.798	0.047	0.142	0.364

Table 19. Descriptive Statistics

5 Results

5.1 MiFID II and Analyst Coverage

The results from Table 20 show that MiFID II has led to a reduction in analyst coverage; this confirms the findings of previous studies (e.g., Guo and Mota 2021). The magnitude of the MiFID II-induced coverage loss amounts to around 0.5 analysts (see Table 20, Model [3]). This magnitude is also comparable to the results of B. Fang et al. (2020) and Guo and Mota (2021). Although this first analysis step does not lead to new findings, it indicates that the sample and variables are suitable for investigating MiFID II-related effects.

Variables	[1]	[2]	[3]	[4]
	N_ANALYSTS			
	Full Sample		PSM Sample	
POST_TREAT	-0.400***	-0.666***	-0.530***	-0.717***
	(-7.21)	(-8.93)	(-7.31)	(-8.03)
POST_TREAT_SMALL		0.473***		0.629***
		(4.51)		(4.69)
POST_SMALL		0.516***		0.558***
		(6.38)		(5.59)
ln(MARKETCAP)	0.880***	0.883***	0.916***	0.911***
	(15.99)	(16.12)	(12.25)	(12.21)
ln(TOTAL_ASSETS)	0.785***	0.744***	0.835***	0.821***
	(9.65)	(9.24)	(8.06)	(8.08)
PPE_RATIO	0.388	0.179	0.127	-0.0870
	(1.22)	(0.57)	(0.29)	(-0.20)
DEBT_RATIO	-0.122	-0.167	0.0530	-0.0423
	(-0.99)	(-1.37)	(0.31)	(-0.25)
MTB	-0.00490	-0.00511	-0.0112**	-0.0116**
	(-1.42)	(-1.47)	(-2.46)	(-2.51)
LOSS	-0.0652	-0.0570	-0.0651	-0.0450
	(-1.30)	(-1.16)	(-1.00)	(-0.70)
ROA	-0.744***	-0.756***	-0.685***	-0.728***
	(-4.89)	(-5.14)	(-3.25)	(-3.67)
TOTAL_RETURN	0.257***	0.270***	0.290***	0.310***
	(6.72)	(7.10)	(5.47)	(5.89)
VOLATILITY	-0.756***	-0.789***	-0.979***	-0.994***
	(-4.56)	(-4.82)	(-4.04)	(-4.17)
TRADING_VOLUME	0.482***	0.472***	0.589***	0.581***
	(7.15)	(7.05)	(6.06)	(6.02)
EPS_DISPERSION	-0.0774***	-0.0766***	-0.104***	-0.107***
	(-3.72)	(-3.75)	(-3.43)	(-3.60)
INTERCEPT	-3.256***	-2.975***	-3.576***	-3.411***
	(-5.51)	(-5.13)	(-4.44)	(-4.37)
N	18,316	18,316	11,808	11,808
Adj. R^2	0.966	0.967	0.962	0.963
Fixed Effects	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Clustering	Firm	Firm	Firm	Firm

t-statistics in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 20. Regression Results – Analyst Coverage and MiFID II

To understand the role of company size we follow Lang et al. (2021) and define SMALL companies as those companies belonging to the lower tercile in terms of total assets in 2016. It appears that small companies are not or only marginally affected by the coverage loss. This can be seen from the positive interaction term POST_TREAT_SMALL, which almost offsets the negative regression coefficient from POST_TREAT (see Table 20, Models [2] and [4]). This confirms the findings of Guo and Mota (2021), Lang et al. (2021), and Amzallag et al. (2021), who were all unable to detect a coverage loss in small companies. The contrary result of B. Fang et al. (2020) cannot be confirmed. All results are robust to the use of the full or the PSM sample.

5.2 Analyst Coverage Loss and Media Coverage

The results for Equation (7) that are tabulated in Table 21 provide an initial supporting indication for **H1**. If any loss in analyst coverage (ANALYST_LOSS) is considered as a treatment (as in Models [1] and [2]), there is no significant negative effect on media coverage. For full articles, even a slightly significant positive effect on media coverage is observable. However, if the treatment is restricted to more severe analyst coverage losses of at least 20% compared to the pre-MiFID era (ANALYST_STRONG_LOSS), the effects on media coverage (full and flash articles) become highly significant. This is consistent with econometric intuition, as a stronger decline in analyst coverage affects the information environment for journalists more than only a marginal decline. In addition, the decline in full articles is more pronounced (around 23% loss in media coverage) than in flash articles (around 14% loss). This is also in line with the expectations from **H3**.

Variables	[1]	[2]	[3]	[4]
	ln(1+N_FULL_ARTICLES)	ln(1+N_FLASH_ARTICLES)	ln(1+N_FULL_ARTICLES)	ln(1+N_FLASH_ARTICLES)
	EU Sample with stable pre-MiFID II Analyst Coverage			
POST_ANALYST_LOSS	0.106**	-0.0500		
	(1.99)	(-1.10)		
POST_ANALYST_STRONG_LOSS			-0.270***	-0.147***
			(-4.49)	(-2.65)
ln(MARKETCAP)	0.0821	0.0444	0.0398	0.0320
	(1.57)	(1.05)	(0.78)	(0.76)
ln(TOTAL_ASSETS)	0.0999	0.0498	0.0806	0.0461
	(1.21)	(0.80)	(1.00)	(0.74)
PPE_RATIO	-0.293	0.821**	-0.260	0.842**
	(-0.93)	(2.08)	(-0.82)	(2.14)
DEBT_RATIO	-0.157	-0.208	-0.107	-0.184
	(-1.11)	(-1.07)	(-0.76)	(-0.96)
MTB	-0.000547	-0.00106	0.000179	-0.000812
	(-0.19)	(-0.40)	(0.06)	(-0.31)
LOSS	0.149***	0.148***	0.151***	0.150***
	(2.86)	(3.04)	(2.92)	(3.08)
ROA	0.0127	0.350*	-0.00769	0.351*
	(0.07)	(1.85)	(-0.04)	(1.83)
TOTAL_RETURN	0.0157	-0.0522	0.0181	-0.0548
	(0.43)	(-1.49)	(0.50)	(-1.57)
VOLATILITY	0.0121	-0.0852	0.143	-0.0300
	(0.07)	(-0.59)	(0.85)	(-0.21)
TRADING_VOLUME	0.0722	0.0948	0.0422	0.0834
	(1.16)	(1.47)	(0.66)	(1.27)
EPS_DISPERSION	-0.00606	-0.00170	-0.00159	-0.00182
	(-0.26)	(-0.12)	(-0.07)	(-0.14)
INTERCEPT	1.416**	2.860***	1.858***	2.950***
	(2.32)	(6.97)	(3.10)	(7.22)
N	3,408	3,408	3,408	3,408
Adj. R^2	0.906	0.900	0.907	0.901
Fixed Effects	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Clustering	Firm	Firm	Firm	Firm

t-statistics in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 21. Regression Results – Loss of Analyst and Media Coverage

5.3 MiFID II and Media Coverage

To link the previously identified effects of media coverage to the MiFID II regulation, media coverage is incorporated as a dependent variable and the introduction of MiFID II is incorporated as a treatment in the difference-in-differences design, as already done in Section 5.1. For the exact specification of the regression equation, see Equation (8). As shown in Models [1] and [2] of Table 22, the media coverage of full articles is reduced on the full (PSM) sample by approximately 12% (7%) after MiFID II implementation. This confirms **H1** and shows that under the MiFID II regulation, journalists contribute significantly less to the dissemination of information among investors in order to reduce information asymmetries. However, MiFID II has no significant impact on media coverage from flash articles. This, in turn, confirms **H3**, according to which flash articles are less affected than full articles because less input is needed for their production, which journalists could obtain from sell-side analysts.

Variables	[1]	[2]	[3]	[4]	[5]
	ln(1+N_FULL_ARTICLES)			ln(1+N_FLASH_ARTICLES)	
	Full Sample	PSM Sample	Full Sample	Full Sample	PSM Sample
POST_TREAT	-0.124***	-0.0679***	-0.100***	0.00894	0.0114
	(-7.18)	(-3.14)	(-4.48)	(0.57)	(0.68)
POST_TREAT_SMALL			-0.0325		
			(-0.90)		
POST_SMALL			-0.0759***		
			(-4.38)		
ln(MARKETCAP)	0.0897***	0.0988***	0.0888***	0.106***	0.0999***
	(5.37)	(4.56)	(5.29)	(6.75)	(5.58)
ln(TOTAL_ASSETS)	0.145***	0.108***	0.150***	0.132***	0.137***
	(6.22)	(3.39)	(6.42)	(5.97)	(5.50)
PPE_RATIO	-0.316***	-0.324**	-0.291***	0.124	0.215
	(-2.87)	(-2.42)	(-2.67)	(1.13)	(1.60)
DEBT_RATIO	-0.0781**	-0.154***	-0.0730*	-0.0281	-0.0478
	(-2.05)	(-3.50)	(-1.91)	(-0.71)	(-1.13)
MTB	-0.000438	-0.00178	-0.000393	-0.00108	-0.000326
	(-0.47)	(-1.40)	(-0.42)	(-1.25)	(-0.29)
LOSS	0.0358**	0.0456**	0.0349**	0.0312*	0.0381**
	(2.11)	(2.12)	(2.06)	(1.86)	(1.97)
ROA	-0.0998*	-0.187**	-0.0978*	-0.104*	-0.0399
	(-1.85)	(-2.26)	(-1.80)	(-1.82)	(-0.50)
TOTAL_RETURN	0.0155	0.0133	0.0140	-0.0169	-0.0221
	(1.26)	(0.82)	(1.13)	(-1.36)	(-1.44)
VOLATILITY	0.249***	0.318***	0.251***	0.0323	0.00907
	(5.35)	(4.97)	(5.42)	(0.74)	(0.16)
TRADING_VOLUME	0.122***	0.134***	0.125***	0.0929***	0.133***
	(5.65)	(4.25)	(5.74)	(4.10)	(4.53)
EPS_DISPERSION	0.00718	0.0117	0.00708	-0.00469	-0.00346
	(0.97)	(1.17)	(0.96)	(-0.92)	(-0.57)
INTERCEPT	2.026***	2.371***	1.997***	2.486***	2.608***
	(12.46)	(10.45)	(12.27)	(14.55)	(12.29)
N	18,316	11,808	18,316	18,316	11,808
Adj. R^2	0.943	0.929	0.943	0.925	0.919
Fixed Effects	Firm & Year	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Clustering	Firm	Firm	Firm	Firm	Firm

t-statistics in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 22. Regression Results – Media Coverage and MiFID II

Since larger companies are more affected by the loss of analysts than small companies are, as the analysis in Section 5.1 depicts, media coverage is also expected to decline more in this company-size segment. However, the regression coefficient of the triple interaction term POST_TREAT_SMALL in Model [3] of Table 22 shows no significant difference in MiFID II-induced media coverage loss between large and small companies. This result does not change when using the PSM sample instead.[6] Therefore, based on the conducted analysis, **H2** cannot be confirmed.

[6] For the stake of brevity, the regression model is not tabulated in Table 22 but is available upon request.

Splitting the media coverage into the EA period and non-earnings announcement (NEA) period shows a stronger reduction during the EA period. Thus, the loss in the EA period is approximately 12%, while that in the remaining period is only about 9%. This has resulted in a shift in the proportion of full articles published during EAs to all full articles. This ratio amounted to 35.1% for the EEA companies before the introduction of MiFID II and was later reduced significantly by MiFID II by approximately four percentage points, as shown in Model [3] of Table 23.[7] This finding confirms **H4** and contradicts the prior findings of Guest and Kim (2022), who found a shift in media coverage toward the EA date. The results remain robust when applying the PSM sample.[8] Our finding is particularly relevant because media coverage has been shown to reduce information asymmetries especially during EAs (Bushee et al. 2010).

[7] The proportion of full articles that is published during the earnings announcements is only calculated if the media coverage (N_FULL_ARTICLES) in the respective firm-year amounts to at least one article. For this reason, the number of observations of model [3] in Table 23 lies below the other models.

[8] For the stake of brevity, the regression model is not tabulated in Table 23 but is available upon request.

Variables	[1]	[2]	[3]
	ln(1+N_FULL_ARTICLES)		PROP_FULL_ ARTICLES_EA
	EA Full Sample	NEA Full Sample	Full Sample
POST_TREAT	-0.130***	-0.0918***	-0.0411***
	(-8.71)	(-4.95)	(-6.81)
ln(MARKETCAP)	0.0811***	0.0953***	0.000360
	(5.20)	(5.40)	(0.08)
ln(TOTAL_ASSETS)	0.0569**	0.150***	-0.0158**
	(2.32)	(6.04)	(-2.57)
PPE_RATIO	-0.0475	-0.391***	0.0796***
	(-0.54)	(-3.18)	(2.58)
DEBT_RATIO	-0.0342	-0.0738*	-0.000171
	(-1.04)	(-1.74)	(-0.02)
MTB	-0.000973	0.0000199	-0.000207
	(-1.10)	(0.02)	(-0.81)
LOSS	0.0375**	0.0408**	-0.00481
	(2.23)	(2.16)	(-0.84)
ROA	-0.0281	-0.107*	0.00984
	(-0.50)	(-1.88)	(0.67)
TOTAL_RETURN	0.00782	0.0264*	-0.00413
	(0.71)	(1.94)	(-1.05)
VOLATILITY	0.164***	0.278***	-0.0202
	(3.65)	(5.41)	(-1.50)
TRADING_VOLUME	0.0719***	0.142***	-0.0165***
	(3.70)	(6.16)	(-3.11)
EPS_DISPERSION	0.0121*	0.0105	0.00117
	(1.83)	(1.30)	(0.50)
INTERCEPT	1.625***	1.605***	0.426***
	(10.31)	(9.10)	(8.71)
N	18,316	18,316	15,669
Adj. R^2	0.934	0.929	0.366
Fixed Effects	Firm & Year	Firm & Year	Firm & Year
Clustering	Firm	Firm	Firm

t-statistics in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 23. *Regression Results – Media Coverage and MiFID II During Earnings Announcements*

5.4 Effect Channel and Robustness Checks

Taking the collective results from Sections 5.2 and 5.3, it seems plausible that the reduced media coverage is a causal effect from MiFID II's research unbundling that is transmitted via a reduction in analyst coverage. However, at this stage, it cannot be ruled out that one of the many other measures required by MiFID II causes this effect. To analyze the effect channel in more detail, Equation (8) is estimated again by using a sample in which the within-firm variation in analyst coverage has been reduced. All companies that experienced a change in analyst coverage by more than 20% compared to the pre-MiFID II time period are removed from this analysis. The results of this analysis are listed in Table 24. The absolute value of POST_TREAT's regression coefficient is strongly reduced compared to that in the original analysis in Table 22. The interaction term is barely insignificant[9] on the full sample and clearly insignificant if propensity score matching is applied. The disappearance of the effect when the variation in analyst coverage is reduced provides further evidence that the decrease in media coverage is causally linked to MiFID II's research unbundling. This regulation reaches journalists via the impact channel of reduced analyst coverage. If the observed loss of media coverage would be due to another reason (e.g., another regulation required by MiFID II), the disappearance of the effect as soon as the variation in analyst coverage is reduced would not be reasonable.

Variables	[1]	[2]
	ln(1+N_FULL_ARTICLES)	
	Reduced Sample	Reduced PSM Sample
POST_TREAT	-0.0454*	-0.0266
	(-1.95)	(-1.03)
ln(MARKETCAP)	-0.00243	0.0306
	(-0.11)	(1.16)
ln(TOTAL_ASSETS)	0.142***	0.130***
	(3.92)	(3.00)
PPE_RATIO	-0.391**	-0.380**
	(-2.41)	(-2.17)
DEBT_RATIO	-0.111**	-0.143**
	(-2.54)	(-2.52)
MTB	-0.00104	-0.00293**
	(-1.09)	(-2.09)
LOSS	0.0720***	0.0909***
	(3.10)	(3.41)
ROA	0.000644	0.0585
	(0.01)	(0.43)
TOTAL_RETURN	-0.00546	-0.0126
	(-0.32)	(-0.59)
VOLATILITY	0.245***	0.335***
	(3.27)	(3.89)
TRADING_VOLUME	0.129***	0.166***
	(3.35)	(3.61)
EPS_DISPERSION	0.000705	-0.00488
	(0.06)	(-0.34)
INTERCEPT	3.051***	2.892***
	(10.51)	(8.43)
N	9,924	8,144
Adj. R^2	0.943	0.923
Fixed Effects	Firm & Year	Firm & Year
Clustering	Firm	Firm
t-statistics in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$		

Table 24. Regression Results – Reduced Variation in Analyst Coverage Changes

The difference-in-differences approach chosen here is subject to the parallel trend assumption. This assumption states that the control and treatment groups develop in parallel if the treatment does not occur (Angrist and Pischke 2008). To verify this, a placebo test is performed on all the regression analyses from Table 22 and Table 23 where

[9] At a 5% level of significance.

the impact of MiFID II on media output is analyzed (see Table 25). For the placebo test, only the pre-treatment period is considered, and January 1, 2017 is chosen as the placebo treatment date. With respect to the analyses from Table 22 no significant effect of the placebo treatment can be observed for the relevant interaction terms if the PSM sample is used. For the full sample, a significant negative effect on the media coverage is observable. However, this effect amounts to less than 50% of the effect measured in the original analysis and might be attributed to a pre-trend. For the placebo testing regarding the analysis on **H4** (EA vs. NEA) no significant response on the placebo treatment is measured for model [1] and [2]; however, for model [3] analyzing the impact on PROP_FULL_ ARTICLES_EA, a significant regression coefficient for the interaction term POST_TREAT is observable. The effect is reduced by around 50% compared to the effect measured in the actual analysis and may also be attributable to a pre-trend.

Table	Model	Interaction Term	Sample	Coefficient	t-value
Table 22	[1]	POST_TREAT	Full	-0.0564***	-2.88
Table 22	[2]	POST_TREAT	PSM	-0.0417	-1.59
Table 22	[3]	POST_TREAT_SMALL	Full	0.0402	1.13
Table 22	[4]	POST_TREAT	Full	-0.0206	-1.57
Table 22	[5]	POST_TREAT	PSM	-0.000008	-0.00
Table 23	[1]	POST_TREAT	EA Full	0.000251	0.01
Table 23	[2]	POST_TREAT	NEA Full	0.0332	1.59
Table 23	[3]	POST_TREAT	Full	-0.0228***	-2.73

* p < 0.1, ** p < 0.05, *** p < 0.01

Table 25. Placebo Testing

6 Discussion

The paper provides evidence that MiFID II has led to a reduction in media and analyst coverage. This has worsened the overall information environment, as media and analysts have contributed less to reducing information asymmetries. The resulting principal–agent problems can increase agency costs and decrease economic welfare. The presented findings offer an explanation for why MiFID II has led to a reduction in liquidity (B. Fang et al. 2020; Lang et al. 2021).

The present paper builds on the literature examining the effects of MiFID II's research unbundling (e.g., Guo and Mota 2021). In contrast to those papers, it examines not only the direct effects on analysts but also secondary effects on other information intermediaries (namely journalists). This paper also complements the literature strand exploring the interaction between financial analysts and media. In particular, it builds on the findings of Guest and Kim (2022) and examines their findings, which are based on a more than 10-year-old sample of brokerage mergers in a contemporary setting with significantly changed underlying conditions. The main findings of Guest and Kim

(2022) are confirmed by this paper. However, unlike Guest and Kim (2022) finding, this paper finds that the loss of media coverage primarily occurs during EAs and to a lesser extent outside of these events. In addition, the media coverage is measured much more broadly in this paper compared to Guest and Kim (2022), since the full set of RavenPack's sources, which are more than 25,000 outlets, are used and not only the outlets from Dow Jones.

This study has its limitations. Despite RavenPack's comprehensive news database, there still exist other media sources that are not covered. The same applies to the analyst database I/B/E/S, which only contains analyst opinions that have been (deliberately) published by the analysts or brokerage houses (Amiram et al. 2018). Moreover, due to missing data in the control variables, numerous companies had to be excluded from the analysis, as shown in Table 18. The placebo test is positive in two cases. This might question the parallel trend assumption of the underlying analysis. However, in the MiFID II setting, the new rules were available to the public long before the actual implementation which increases the chances of pre-trends. The results of the placebo test are likely to be explained by a pre-trend which would in fact lead to an underestimation of the measured effects. This provides confidence in the validity of the results despite the two positive placebo treatments. Furthermore, for the main hypothesis of this study (**H1**) the positive finding from the placebo test is not confirmed on the PSM sample which highlights the importance of PSM.

Methodologically, a central problem lies in the identification of the treatment. From a legal point of view, the jurisdiction of the asset manager and not of the covered company is relevant for the applicability of MiFID II. However, since this information is not available and there is no link to the media coverage, the strategy employed by Guo and Mota (2021) is used, and the treatment and control groups are assigned according to the company's headquarter location, about which analysts and journalists write analyses and texts. It should be noted, however, that this approach may lead to an underestimation of the effect sizes, as US companies will also be affected to some extent by MiFID II (Allen 2019; Guo and Mota 2021).

This paper has implications for practitioners and researchers. Regulators can use the findings to review and, if necessary, adjust regulation if it has unintended effects that negatively impact capital markets. This is particularly relevant in this case because even though MiFID II is intended to strengthen transparency, the decrease in media coverage may have a negative impact on information transparency for small investors. As the paper shows how far-reaching the effects of the regulatory measures can be, it highlights the importance of holistically researching the effects on information intermediaries. The example of journalists and sell-side analysts demonstrates that strong interrelations exist between information intermediaries.

7 Conclusion

MiFID II has comprehensively changed the market of sell-side analysts. Before the actual introduction of MiFID II, many experts already anticipated its impact on analysts and the decrease in their coverage (Preece 2017). However, as analysts do not operate in isolation but as part of the complex capital market ecosystem, the regulation could also impact many other market participants. This paper shows that research unbundling has also had unintended effects on the provision of information by journalists. It turns out that their media coverage decreased by between 12% to 7% due to MiFID II. This is in clear contradiction to the objectives of MiFID II, according to which transparency and investor protection should be increased. However, it is precisely the private investors, who have no or only limited access to exclusive information, who are dependent on journalistic content as a source of information (L. Fang and Peress 2009).

Subsequent research should evaluate whether MiFID II has also changed the content of news articles, as the quantitative aspect of news articles was analyzed extensively, but no attention was paid to the content of these articles. In addition, future research could examine a MiFID II-induced change in the magnitude of capital market reactions following the publication of news. Similar to the findings on analyst coverage (e.g., B. Fang et al. 2020), the reduction of coverage could lead to stronger capital market reactions on individual news articles.

Appendix – Variable Definition

Variable	Definition	Source
N_ANALYSTS	Number of analysts that forecasted the EPS during the respective year	I/B/E/S
N_FULL_ARTICLES	Number of full articles about the company during the respective year	RavenPack
N_FLASH_ARTICLES	Number of flash articles about the company during the respective year	RavenPack
N_FULL_ARTICLES_EA	Number of full articles about the company that are posted during the three days surrounding the earnings announcements	RavenPack
N_FULL_ARTICLES_NEA	Number of full articles about the company that are posted outside the three days surrounding the earnings announcements	RavenPack
PROP_FULL_ ARTICLES_EA	N_FULL_ARTICLES_EA divided by N_FULL_ARTICLES for the respective firm-year (the variable is only calculated for N_FULL_ARTICLES > 0)	RavenPack
MARKETCAP	Market capitalization of the company measured in USD million (MV)	Datastream
TOTAL_ASSETS	Total assets (WC02999) of the company measured in USD million	Datastream
PPE_RATIO	Book value of property, plant, and equipment (WC02501) divided by total assets	Datastream
DEBT_RATIO	Total debt divided by total capital (WC08221)	Datastream
MTB	Market value of equity divided by book value of equity (PTBV)	Datastream
LOSS	Indicator variable that takes the value 1 if the net income (WC01706) is negative and 0 otherwise	Datastream
ROA	Return on assets (WC08326)	Datastream
TOTAL_RETURN	The annual total return of an investor holding the asset for the respective year (WC08801)	Datastream
VOLATILITY	Annualized volatility of daily returns	Refinitiv Workspace
TRADING_VOLUME	Annual trading volume divided by common shares	Refinitiv Workspace
EPS_DISPERSION	Standard deviation of EPS forecasts divided by the median EPS forecast. With two or fewer EPS forecasts for the respective year, the mean industry (TRBC business sector) dispersion is used.	I/B/E/S; Refinitiv Workspace
POST	Indicator variable that takes the value of 1 for the years 2018 and onwards (post-MiFID II era) and 0 otherwise	
TREAT	Indicator variable based on headquarter location, with 1 for EEA companies and 0 for US companies	Refinitiv Workspace
ANALYST_LOSS	Indicator variable that is set to 1 if the average N_Analyst during the post-MiFID era is below the average N_Analyst in the pre-MiFID II era and 0 otherwise	I/B/E/S
ANALYST_STRONG_LOSS	Indicator variable that is set to 1 if analyst coverage during the post-MiFID era is at least 20% below the coverage in the pre-MiFID II era and 0 otherwise	I/B/E/S
SMALL	Indicator variable that is set to 1 if TOTAL_ASSET (USD) in 2016 is below the lower tercile	Datastream

Table 26. Variable Definition

II.4. Signaling Digital Transformation

The Role of CDOs in Signaling Digital Transformation Endeavors: An Analysis of Firms' External Communication Tools

Authors: Dennis R. Metzler, Steffen Bankamp, Jan Muntermann and Matthias Palmer

Outlet: *Proceedings of the 42nd International Conference on Information Systems, Austin, United States, 2021.*

Abstract: As part of their digital transformation, firms increasingly appoint Chief Digital Officers (CDOs). Existing research suggests that CDOs are appointed to drive and coordinate digital transformation activities and communicate digital transformation-related topics to stakeholders. However, the specific role of the CDO as a mediator between a firm and its external stakeholders, such as investors, remains unclear. Relying on signaling theory, we investigate whether CDO presence impacts digital transformation-related signaling in firms external communication tools. Indeed, our results show a strong positive association between CDO presence and the volume of digital transformation-related signals. Therefore, it can be assumed that CDO presence has the potential to contribute to reducing digital transformation-related information asymmetries between firms and external stakeholders. However, since our results show that less regulated communication tools are more likely to be used for digital transformation-related signaling than highly regulated ones, the reliability of such signals remains questionable.

Keywords: Digital Transformation, Chief Digital Officer, External Corporate Communication, Signaling Theory, Quantitative Research, Natural Language Processing

1 Introduction

With rapid advancements in the development and improvement of digital technologies, firms must increasingly address the challenges of digitalization. Thereby, digitalization and associated technological innovations lead to disruptions within industries and markets and to rapidly changing organizational environments (Bharadwaj et al. 2013; Verhoef et al. 2021). To stay competitive in an increasingly digitalized society, firms need to evolve and adapt to the changing business landscape, making digital transformation crucial for firms to survive and remain competitive (Bharadwaj et al. 2013; Firk et al. 2021; Vial 2019). In recent years, an increasing number of firms have recognized the need for digital transformation and its potential opportunities. In that regard, firms increasingly consider digital transformation a critical success factor and invest in new technologies and associated capabilities (Sebastian et al. 2017).

As digital transformation becomes a high-level imperative for firms and their stakeholders, it has turned into a high precedence concern on the leadership level (Hess et al. 2016). The leadership, comprising the board of directors and the rest of the top management team, is vital to a firm's digital transformation. It is responsible for driving and coordinating the strategic direction of an organization, including the decision on how to address digital transformation (Luciano et al. 2020). In addition, the top management team is responsible for communicating digital transformation-related topics with important stakeholders, such as investors (e.g., Singh and Hess 2017). In that regard, an increasing number of firms appoint the position of the Chief Digital Officer (CDO) to the top management team as a centralized digital transformation responsibility with the aim to drive and coordinate digital transformation and to communicate digital transformation-related topics with stakeholders (R. Grossman and Rich 2012; Péladeau et al. 2017; Singh and Hess 2017; Kunisch et al. 2022; Singh et al. 2020).

Existing research on CDOs writes from different perspectives. In that regard, Kessel and Graf-Vlachy (2021) found that CDO-related research can primarily be distinguished in three different research streams: (1) Antecedents of CDO presence, (2) The CDO in the organization, and (3) Consequences of CDO presence. Whereas research on antecedents of CDO presence and the CDO in the organization is already advanced, research on the consequences of CDO presence is somewhat underrepresented in the existing literature (Kessel and Graf-Vlachy 2021). Thereby, most of the existing research concerning the consequences of CDO presence deals with the impact of CDOs on innovation performance (e.g., Leonhardt et al. 2018; Reck and Fliaster 2018, 2019) or on financial performance (e.g., Zhan and Mu 2016; Drechsler et al. 2019; Berman et al. 2020; Firk et al. 2021). However, although Singh and Hess (2017) found that an appointed CDO is responsible for communicating digital transformation-related topics with stakeholders, the specific role of the CDO as a mediator between a firm and its external stakeholders, such as investors, is still scarcely investigated. In that regard, the reduction of potential information asymmetries between firms and external stakeholders could be a potential side effect of CDO presence. From a signaling perspective, Drechsler et al. (2019) show that firms use the announcement of CDO appointments as a form of strategic signaling to investors. However, since digital transformation activities of firms are bound to risk and uncertainty (e.g., Hess et al. 2016; Sebastian et al. 2017; Moker et al. 2020), related information are highly relevant to evaluate the future prospects of a firm. Therefore, firms need to further send digital transformation-related signals to external stakeholders to reduce potential information asymmetries. Overall, we assume that digital transformation-related signaling does not only include the announcements of CDO appointments.

Research on digital transformation-related signaling is still rare. It especially remains unclear whether those firms appointing a CDO are more likely to conduct digital transformation-related signaling, especially in their external communication tools. If so, it could be assumed that CDO presence can be seen as an indicator for better digital transformation-related signaling and that CDO presence has the potential to reduce digital transformation-related information asymmetries between firms and external stakeholders. Due to the high relevance of digital transformation for the future competitiveness of firms and the resulting high relevance of digital transformation-related information for its stakeholders, especially investors, this research gap should be closed. We approach this research gap by analyzing whether the presence of a CDO can be associated with a higher volume of digital transformation-related signaling in firms' external communication tools. In addition, in order to investigate the reliability of digital transformation-related signaling, we further analyze potential differences between communication tools with different degrees of regulation. Against this background, we formulate the following research questions (*RQs*):

RQ1: How does CDO presence impact the volume of digital transformation-related signals in external communication tools?

RQ2: How does the volume of digital transformation-related signals differ across communication tools with different degrees of regulation?

To answer these research questions, we derive two hypotheses from the literature and analyze the relationship between CDO presence in a firm's top management team and digital transformation-related signaling in external communication tools. Thereby, the volume of these theme-specific signals is measured by the relative frequency of digital transformation-related sentences within the main external communication tools firms use to communicate with external stakeholders and reduce potential information asymmetries. To calculate the frequency of digital transformation-related sentences, we used the dictionary of digital terms developed by W. Chen and Srinivasan (2019), which we further extended by keywords related to digital technologies and digitalization in general. Our study focuses on the constituents of the S&P 500 equity index from 2007 to 2020. Based on insights from existing research on digital transformation and CDOs, we assume that firms appointing a CDO to their top management team pay increased attention to digital transformation activities. In addition, the appointed CDO should further drive digital transformation and digital transformation-related communication and thereby further bring digital transformation to a firm's focus. Overall, this should result in an increase in the volume of digital transformation-related signals. In that regard, we distinguish between highly regulated communication tools (10-K reports) and less regulated communication tools (conference calls). Both communication tools are highly relevant, but they differ significantly in their degree of regulation and subsequently in their reliability, which may impact how firms use them to communicate digital transformation-related information and how relevant they are for external stakeholders, especially investors.

We contribute to the existing literature concerning the consequences of CDO presence in manifold ways. Our study holds important implications for firms deciding whether to appoint a CDO or not and stakeholders deciding which firms are more inclined to signal digital transformation-related activities and where to search for digital transformation-related signals. Our results show that firms with a CDO in their top management team are accompanied by a significantly higher volume of digital transformation-related signals in their external communication. However, we further show that the increase in the volume of digital transformation-related signals in less regulated communication tools is significantly higher than in highly regulated communication tools which questions the reliability of such signals.

To provide sound theoretical foundations and gain valuable insights regarding our research questions, this paper is structured as follows: Starting with the theoretical foundations, we introduce the role of the CDO in the digital transformation journey as well as the role of signaling in corporate communication. Secondly, we introduce the methodological foundation of the conducted study. Thirdly, we present the findings of our analysis. Fourthly, in the context of a discussion, the limitations of our study and implications for future research and practice are presented. Finally, the conclusion summarizes the most important findings.

2 Theoretical Foundations

2.1 The CDO as the Centralized Digital Transformation Responsibility in Firms

The emergence of new digital technologies has a transformational impact on today's society. In a business context, digital technologies can reconfigure the way firms operate their business, communicate with stakeholders (e.g., customers and partners), and compete within markets (Bharadwaj et al. 2013; Hess et al. 2016). The changes that digital technologies bring to a firms' business model, resulting in changed products, the automation of processes, or changed organizational structures, can be described as digital transformation (Hess et al. 2016). Firms need to undergo a digital transformation to stay competitive in an increasingly digitalized market environment and thereby adapt their current business models, organizational structures, strategy, and internal culture (e.g., Matt et al. 2015; Eden et al. 2019; Metzler and Muntermann 2020). For this reason, the process of digital transformation can be seen as one of the most relevant topics on the agenda of executives across industries.

Existing research indicates that a firm's leadership team and especially its top management team play an important role in the strategic change processes of firms, such as the digital transformation (Singh et al. 2020). Since digital transformation involves a fundamental transformation of the entire organization, including the need for adapting mindsets and skillsets, leadership is a crucial factor in the process of digital transformation (Westerman et al. 2014). In order to adapt the top management team for the digital era and subsequently drive digital transformation, an increasing number of firms appoint new technology-related C-level roles to the top management team. This, for example, includes the Chief Information Officer, Chief Innovation Officer, Chief Data Officer, Chief Strategy Officer, and the Chief Digital Officer. Chief Information Officers are in charge of IT support and IT deployment, Chief Innovation Officers are in charge of corporate in general without a specific digitalization focus, Chief Data Officers are responsible for the data management and data analytics, and Chief Strategy Officers are responsible for managing and executing strategy processes. Finally, the Chief Digital Officer can be described as the key position of highest responsibility for digital transformation in firms. The CDO is responsible for driving digital transformation activities, digital mobilizing the entire firm, initiating firm-wide collaboration, and communicating digital transformation-related topics with stakeholders (e.g., Singh and Hess 2017).

Not all firms appoint a CDO to the top management team to drive digital transformation. For example, various management boards believe that an already existent CIO is sufficient to fulfill this task. However, in that regard, Singh and Hess (2017) mention that, due to the complexity of digital transformation, it is challenging for a CIO to manage the digital transformation in addition to the original responsibilities of the CIO. Therefore, a CIO might not be the best choice for managing a firm's digital transformation. Other opportunities include, but are not limited to, giving the digital transformation responsibility to the CEO (Hess et al. 2016) or divisional or functional heads (Björkdahl 2020). Overall, existing research does not find a consensus on whether the appointment of a CDO to the top management team is an adequate decision concerning digital transformation issues. Therefore, it remains unclear whether the appointment of a CDO is an essential success factor in the process of digital transformation (e.g., Leonhardt et al. 2018). However, when appointing a CDO to the top management team, it is essential that the CDO and other C-level positions work closely together. For example, the CIO provides the foundation for digital transformation by delivering the necessary agile IT capabilities for more flexibility and digital innovation (Haffke et al. 2016). Furthermore, the CIO is also responsible for implementing the changes in the infrastructure and platforms. Therefore, it is essential that the CIO and the CDO work closely together while the CIO acts as an IT specialist and the CDO as the digital transformation specialist (Haffke et al. 2016; Singh and Hess 2017). Moreover, as the most senior manager, the CEO needs to back the digital transformation and assure that framing the digital transformation successfully supports the CDO in engaging and inspiring the entire organization, especially middle management. Therefore, also the CEO needs to work closely with the CDO and support the digital vision and activities (Westerman et al. 2014).

The decision to appoint a CDO to the top management team depends on various internal and external factors (Kessel and Graf-Vlachy 2021). Most firms appoint a CDO as a response to realizing that the current top management team lacks managers with appropriate skills. In addition, CDOs are most common in firms with a focus on intangible assets. In firms focusing on tangible assets, CDOs are not that frequently presented (Firk et al. 2021; Kessel and Graf-Vlachy 2021). Another common trigger of appointing a CDO is market competition. In markets with highly digital-savvy competitors, firms appoint CDOs as a reaction to their peers (e.g., Haffke et al. 2016; Singh and Hess 2017; Firk et al. 2021; Kessel and Graf-Vlachy 2021).

Exiting research on CDOs has further dealt with the required characteristics and skill-sets of CDOs. In that regard, it was found that a good CDO needs a mixture of technology-related skills (e.g., general IT competencies), a digital mindset (e.g., a digital visionary spirit), and more general skills (e.g., change management expertise) (Singh and Hess 2017). Additionally, existing literature derived various CDO-typologies regarding their specific role within the leadership team. For example, Singh and Hess (2017) proposed three different CDO types: (1) Entrepreneur CDOs, (2) Digital Evangelist CDOs, and (3) Coordinator CDOs. The Entrepreneur CDO mainly focuses on digital innovation, complementing the existing IT infrastructure and drive innovation by developing, exploring, and exploiting digital technology. The Digital Evangelist CDO focuses on spreading the digital strategy throughout the organization to motivate and inspire employees for the digital transformation. Finally, the Coordinator CDO drives high-level coordination and alignment throughout the organization and creates synergies across the firm. However, all CDO-typologies have in common that they agree on the fundamental idea of implementing a CDO: setting up a position in the top management team that drives and coordinates a firm’s digital transformation journey (e.g., Singh and Hess 2017; Tumbas et al. 2017).

Appointed to a firm’s top management team, a CDO drives and coordinates the digital transformation with the responsibility of formulating an overarching digital transformation strategy and making digital transformation a strategic priority (Haffke et al. 2016; Singh and Hess 2017; Singh et al. 2020; Westerman et al. 2014). This includes introducing new digital technologies, driving a digital culture, and accelerating the digital transformation process (Singh and Hess 2017; Singh et al. 2020). In addition, the CDO is responsible for coordinating digital initiatives and the associated change management within a firm, mediating between different organizational units, working against organizational barriers, and communicating digital transformation-related topics with stakeholders (Singh and Hess 2017; Tumbas et al. 2017, 2018). However, it remains unclear whether these actions are also visible and valuable in the communication with external stakeholders, especially investors. In that regard, CDO presence could be associated with a higher volume of digital transformation-related signals that would, at best, reduce potential information asymmetries between a firm and its external stakeholders.

2.2 Signaling Theory and the Reduction of Information Asymmetries Through External Corporate Communication

Information asymmetry frequently occurs between a firm’s management (possessing more information) and different stakeholder groups, especially investors (possessing less information). In that regard, the principal-agent theory explains contractual relations between parties with mismatched goals in the presence of uncertainty and asymmetric information (Pavlou et al. 2007). The principal (e.g., investor) commissions the agent (e.g., manager) to perform tasks on her or his behalf (e.g., management of the firm). In this case, the agent has more precise information than the principal due to her or his specific role and related activities, making the agent’s assessment more difficult. Situations can arise in which the agent does not act in accordance with the principal’s utility function but only maximizes her or his own utility. S.J. Grossman and Hart (1983) show that this situation can reduce investor's welfare.

In order to reduce information asymmetry and minimize potential welfare losses, firms capitalize on signaling. The so-called signaling theory primarily addresses situations where two different parties have asymmetric information concerning a specific topic (Connelly et al. 2011; Spence 2002). In his seminar work on job market signaling, Spence (1973) shows how job applicants can reduce information asymmetry to hamper the selection ability of prospective employers (Connelly et al. 2011). Generally, signaling theory explains how the party with more information (e.g., management), the sender, chose signals to communicate that information. The other party (e.g., investor), the receiver, should interpret this signal (Connelly et al. 2011). How useful and effective a signal is for a potential receiver is determined by signal reliability (or signal credibility) (Connelly et al. 2011; Davila et al. 2003), which can be described as the extent to which a signal can be perceived as trustworthy. One of the most common signaling tools for firms are external communication tools, including 10-Q-reports, 10-K-reports, and conference calls. These tools include information about financials as well as information about current strategic topics, including digital transformation. Although all these communication tools are highly relevant, they are characterized by a different degree of regulation and standardization. On the one hand, 10-Q-reports and 10-K-reports are documents required by the SEC quarterly (10-Q) or yearly (10-K). These documents contain financial statements, disclosures, internal controls, and management discussions and analyses (SEC 2021). The management has to report all material information, including qualitative information (Cannon et al. 2020). In addition, the 10-K reports are audited by external auditors (SEC 2021). Overall, it can be concluded that 10-K reports are highly regulated and standardized documents for corporate disclosure. The content in these documents can be classified as highly trustworthy. On the other hand, also other less regulated tools are used to communicate with external stakeholders. For example, conference calls are quarterly telephone-based meetings where firms inform investors and analysts about current topics concerning their business development. These conference calls play a unique role, as they take place in connection with the quarterly earnings announcements and thus provide an essential form of corporate disclosure (A.H. Huang et al. 2018). In contrast to 10-K reports, conference calls are not one-sided communication, but company representatives also have to respond spontaneously to questions raised by analysts or others. Thus, compared to highly regulated 10-K reports, conference call transcripts are much less standardized and non-audited documents. The trustworthiness concerning its content, therefore, is not necessarily secured.

Existing research shows that firms use signaling to reduce potential information asymmetries with external stakeholders concerning various topics (e.g., Moker et al. 2020).

Since digital transformation activities of firms are bound to risk and uncertainty (e.g., Hess et al. 2016; Moker et al. 2020; Sebastian et al. 2017) and related information highly relevant to evaluate the future prospects of a firm, external stakeholders, such as investors, try to gather a lot of information in order to reduce potential information asymmetries (Moker et al. 2020). Especially a firms' central corporate communication tools are suitable for stakeholders to look out for visible signals of firms (Moker et al. 2020). In that regard, Brown et al. (2004) showed that conference call activity is negatively related to information asymmetry and Fu et al. (2012) show that information asymmetry is reduced when the frequency of financial reporting increases.

Related research shows that the presence of a chief data officer is associated with a higher frequency of big data-related signaling in annual reports (Kralina 2018). Concerning CDOs, Drechsler et al. (2019) found that firms use public announcements of CDO appointments as strategic signaling to investors. However, it remains unclear whether CDO presence is also associated with a higher quantity of digital transformation-related signaling and whether this information is relevant for external stakeholders with regard to potential information asymmetries.

2.3 Hypothesis Development

Existing research agrees that digital transformation is a highly relevant topic concerning the future competitiveness of firms (e.g., Westerman et al. 2014). In order to drive and coordinate digital transformation activities, firms increasingly appoint CDOs to their top management team (e.g., Singh and Hess 2017; Singh et al. 2020; Tumbas et al. 2017). In that regard, it can be assumed that those firms appointing a CDO to their top management team pay particularly increased attention to digital transformation activities. A high strategic priority of digital transformation, paired with the fact that digital transformation activities are bound to risk and uncertainty, holds the risk of information asymmetries between a firm and its external stakeholders. In order to reduce potential information asymmetries, these firms can increase their digital transformation-related signaling. Since an appointed CDO is responsible for communicating digital transformation-related topics with external stakeholders, it should further be recognizable that CDOs increase the strategic priority of digital transformation in external communication. Overall, these circumstances should be visible in digital transformation-related signaling in firms' external communication tools. Existing research underscores these assumptions. For example, Kralina (2018) shows that the appointment of a special position (i.e., chief data officer) to the top management team can be associated with increased signaling in the area of responsibility of this person (i.e., big data activities) (e.g., Kralina 2018). Based on these assumptions, we derive the following hypothesis 1 (*H1*):

H1: CDO presence can be associated with a higher volume of digital transformation-related signals in firms external communication tools.

As already discussed in the theoretical background, there exist differences between external communication tools. Whereas highly regulated communication tools (i.e., 10-K reports) mainly contain strictly defined content, less regulated tools (i.e., conference calls) include information on current topics where the specific information needs of analysts and other stakeholders can be addressed. In that regard, signal reliability is an important issue. On the one hand, from the argument of signal reliability, regulated communication tools would be more appropriate if firms want to signal that they really engage in digital transformation. Since such communication tools are more trustworthy and reliable, their signals are more useful for their receivers. Thus, if firms really put much effort into digital transformation activities, which implies that their signals are meaningful and provable, they would choose these more reliable communication tools for digital transformation-related signaling. However, on the other hand, if digital transformation is more of a cheap talk; that is the firms like to talk about it but not do any substantially with digital transformation, firms would primarily rely on less regulated communication tools to talk about digital topics. Overall, both, firms that strongly engage in digital transformation, as well as firms for those digital transformation is more of a cheap talk, can engage in digital transformation-related signaling in less regulated communication tools. However, only those firms really engage in digital transformation can also engage in digital transformation-related signaling in highly regulated communication tools. In the end, it can be assumed that the volume of signals differs across different communication tools. Only those firms really engaging in digital transformation can use digital transformation-related signaling in highly regulated communication tools, and only CDOs in such firms can further accelerate this signaling. Based on these assumptions, we further derive the following hypothesis 2 (*H2*):

H2: The impact of CDO presence on the volume of digital transformation-related signals in non-regulated communication tools is higher than in regulated ones.

To test our hypotheses, we measure the volume of digital transformation-related signals in different external communication tools of the analyzed firms. In that regard, we use the relative amount of digital transformation-related sentences in two of the most important external communication tools: (1) 10-K reports and (2) conference calls.

Since the CDO appointment is an endogenous and not a random event, firms make a conscious decision to make a CDO appointment. This endogeneity problem makes it difficult to make statements about the causal effect of CDO appointments since an unobserved third variable and not the CDO appointment itself could drive the results. We aim to minimize this problem by selecting an appropriate research methodology.

The following section describes our methodological approach to test the derived hypothesis and subsequently answer our research questions.

3 Methodological Approach

To analyze the impact of CDO presence within a firm's top management team on the volume of digital transformation-related signals, we conduct an empirical study comprising several sequential steps. In the first step, we utilize natural language processing techniques to calculate the relative frequency of digital transformation-related sentences (DIGITAL RATIO) in a firm's major external communication tools. The sentence-based ratio should prevent the use of numerous topic-specific words in a short section of the text, biasing the results as it could happen with a word-based ratio. However, as part of the validity check, we can confirm that the results of this study do not change when the digital ratio is calculated at the word level. The DIGITAL RATIO serves as our proxy to measure the volume of digital transformation-related signals. It is calculated for the selected firms' 10-K reports and conference calls. Indeed, not every digital transformation-related sentence has to be a conscious and deliberate signal in the sense of signaling theory. For example, it may be the case that certain content (especially in 10-K reports) must be reported due to regulatory requirements. Although this kind of communication can reduce information asymmetries, it would lack the conscious decision of the signaler that is at the heart of signaling theory. Since it is hardly possible to decide which sentences were sent conscious and deliberate in the sense of the theory, we cannot make a differentiation and consider all sentences as signals in the sense of the signaling theory.

To determine the relative frequency of digital transformation-related sentences in these documents, we use the dictionary of digital terms developed by W. Chen and Srinivasan (2019) and extend it with other important digital technology-related and other digitalization-related word groups and keywords. The existing dictionary comprises a selection of relevant digital technology-related word groups (i.e., the word groups Big Data, Cloud, Artificial Intelligence, and Machine Learning) with a selection of relevant keywords for each word group and a selection of other digitalization-related keywords. Since this sample of word groups and keywords does not represent a sufficient universe of digital transformation-related issues, we extend the existing dictionary by adding word-groups concerning other important digital technologies. Thereby, we primarily focus on SMACIT technologies (Sebastian et al. 2017) and add the word groups "Social Media," "Mobile," and "Internet of Things." Furthermore, we extend the existing word groups with similar and alternative words. The final dictionary of digital terms can be found in the appendix. Researchers are invited to use and extend the existing dictionary for future research projects. A sentence is classified as digital transformation-related if it contains at least one entry (word or n-gram) from the applied dictionary. In that regard, we use a search that is not case-sensitive. If relevant, we also consider different wordings (e.g., virtual agent / virtual agents). In the appendix, the words for which we consider different endings are indicted by the wildcard character "*." We calculate the DIGITAL RATIO of a document by dividing the digital transformation-related sentences by the total number of sentences in the document.

In order to determine the extent to which CDO presence affects the volume of digital transformation-related signals, we estimate equation (9), representing a panel regression in which the firms are observed several times during the observation period. This panel structure is particularly suitable for investigating an event's effect (in this case, first-time CDO appointments) on the dependent variable (Wooldridge 2015).

$$\begin{aligned}&\text{DIGIAL RATIO[CC;10 K]}_{\text{t,i}} = \alpha_0 + \alpha_1\text{CDO}_{\text{t-1,i}} + \alpha_2\text{INTANGIBLES}_{\text{t}-1,\text{i}}\\&+\alpha_3\text{MTB}_{\text{t}-1,\text{i}} + \alpha_4 \ln(\text{TOTAL ASSETS})_{\text{t}-1,\text{i}} + \alpha_5\text{ROA}_{\text{t}-1,\text{i}} + \alpha_6\text{LEVERAGE}_{\text{t}-1,\text{i}}\\&+\alpha_7\text{RETURN}_{\text{t}-1,\text{i}} + \alpha_8 \ln(1+\text{DIGITAL M \& A})_{\text{t}-1,\text{i}} + \alpha_9\text{RELATED CxO}_{\text{t}-1,\text{i}}\\&+\alpha_{\text{i}} + \alpha_{\text{t}} + \varepsilon_{\text{t,i}}\end{aligned} \tag{9}$$

The dependent variable DIGITALRATIO is calculated separately with respect to the conference calls [CC] and the annual reports [10-K] on a firm (i) year (t) level. In order to answer RQ2, the digital Ratios will be examined separately for each document type. The main variable of interest is the binary variable CDO which is set to 1 for all firm-year combinations with an acting CDO in the respective year and firm and 0 in all other cases. We further incorporate common control variables into the equation that could influence the volume of digital transformation-related signals. We follow Firk et al. (2021) and use intangible assets (excluding goodwill and scaled by net sales) to assess whether the business model is more focused on knowledge (intangible assets) or on tangible assets (e.g., production of raw materials). We use the market-to-book ratio (MTB) to account for the firm's valuation, the natural logarithm of total assets to account for firm size, return on assets (ROA) to account for profitability, the leverage ratio (LEVERAGE) to account for the capital structure and the annual stock return (RETURN) to account for current stock market performance. In addition, we use the variable DIGITAL M&A to account for the acquisition of digital knowledge through inorganic growth (Hanelt et al. 2021). The variable is calculated by the number of digital M&A transactions the company has conducted as an acquirer during the respective year. We define an M&A transaction as digital if the target's business description or the purpose text of the deal contains at least one entry of the dictionary that is also used for the DIGITAL RATIO. Since it is not only the CDO who could potentially engage in digital-transformation-related signaling, we also consider the board's composition with respect to other technology-related C-level roles as discussed in the theoretical background (section 2.1). The variable RELATED CxO is 1 in all firm-years where the board has a chief information/technology/innovation/data or strategy officer and 0 in all other cases. We further utilize firm fixed effects (α_i) to control for all time-invariant firm characteristics (e.g., industry) (Wooldridge 2015) and year fixed effects (α_t) to account for period-specific characteristics (e.g., increased awareness of the relevance of digitalization activities over time). This comprehensive set of controls reduces the problem of endogeneity of a CDO appointment in our research design. Finally, we use lagged independent variables (lagged by one year) to mitigate the potential problem of reversed causality.

4 Datasets and Descriptive Statistics

To get a basic understanding of the main variables used in this study and to present first interesting insights concerning our data, we present our different datasets and descriptive statistics. As a sample, we use all firms that were part of the U.S. equity index S&P 500 at any time during the period from January 1, 2007 to December 31, 2020. The selection of the S&P 500 allows a relatively broad sample and also good data availability. Since the trend towards appointing CDOs started around 15 years ago (Singh and Hess 2017), the period under review covers the main phase of CDO appointments in firms. The resulting sample includes a total number of 810 firms and thus theoretically 11,340 firm-year observations (810 firms * 14 years). Not all 810 firms existed during the whole period (e.g., due to liquidations and mergers). This reduces the number of observations we consider for our analysis.

For our research approach, we use three distinct datasets. The first dataset comprises information about board positions. Our main variable (CDO), as well as the control variable RELATED CxO, is drawn from this dataset. The process of gathering the CDO data comprises several sequential steps. In the first step, we combine the data of the three databases (1) Boardex, (2) Amadeus, and (3) Crunchbase and extracted all current and former senior executives for those firms included in our sample. Afterward, we identify relevant CDO positions. In that regard, we build on recommendations of existing research (e.g., Kunisch et al. 2022). Therefore, we classify all senior executives with the term “digital” in their role title as potential CDOs. According to Kunisch et al. (2022), this procedure ensures considering those CDOs with similar roles but different role titles. In the next step, we check all resulting potential CDOs and eliminate clear non-CDOs. This, among others, include divisional CDOs, subsidiary CDOs, Chief Data Officers, and CIOs. Finally, to extend our dataset, we also browsed professional websites, online-based executive platforms, firm websites, and press releases with regard to those firms included in our sample. The final sample of CDOs solely contains top management positions responsible for the digital transformation activities within their specific firm. Our approach identifies 213 CDOs across 152 firms of the total 810 firms included in our sample. For the control variable RELATED CxO, we used the role titles from the Boardex and Amadeus databases.

The second dataset includes the textual data used to explore the scope of digital transformation-related content in firms' external communication tools. This dataset includes 10-K reports and transcriptions of conference calls for those firms included in our sample. We extract 10-K reports from the SEC EDGAR database, and the conference calls stem from the Refinitiv Thomson ONE database. We choose these data types as they represent highly relevant external communication tools that give insights into ongoing and completed strategic issues. Both data types aim to reduce potential information asymmetries and therefore contain a vast amount of information that enables a deeper insight into the firms' corporate strategy and, therefore, are suitable for our study (e.g., Bowman 1984; Kloptchenko et al. 2004; Lee and Hong 2014). Whereas 10-K reports are highly standardized annual reports whose publication is legally prescribed, conference calls are carried out several times a year (usually each quarter) to inform investors and analysts about a firm's business development. In contrast to other annual reports, 10-K reports are generally more detailed but lack graphical elements. From the 10-K reports, we extract the text passages for further analysis. We remove any tables and figures. From the conference calls transcripts, we separated the content from the metadata. For the subsequent analysis, we use the conference calls' presentations as well as the Q&A sessions. We remove extremely short sentences of less than 20 characters, as a manual review of the text sections has shown that these are mostly very short statements from conference call participants without meaningful content (e.g., "Ok, thank you."). We use the extracted textual data from both sources to calculate the DIGITAL RATIO as described in the methodology section.

To obtain a first understanding of the data used, we combined both datasets (i.e., CDO information and text in external communication tools) to show how the CDO RATIO and the DIGITAL RATIO evolved over time.

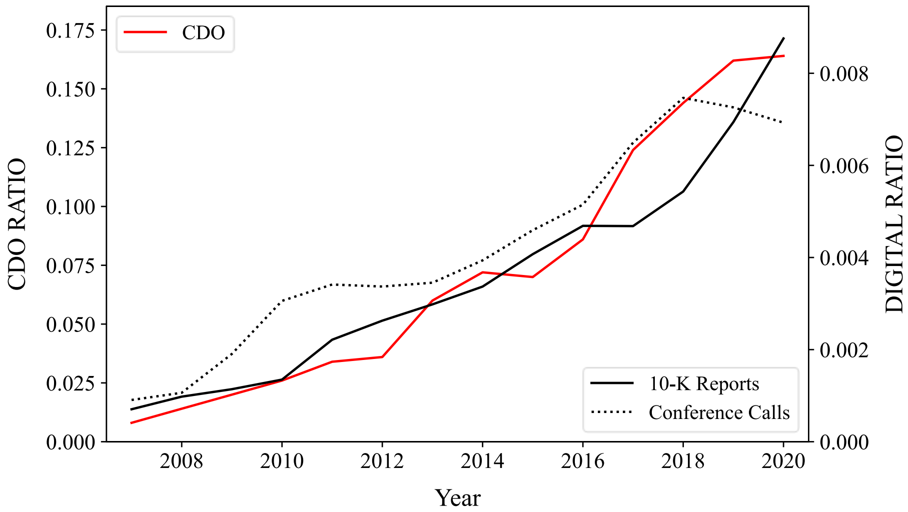

Figure 27. *CDO Presence and Digital Transformation Activities Within S&P 500 Over Time*

As illustrated in Figure 27, the CDO presence across the considered firms has increased strongly during the observation period. In line with existing literature, we can observe that before 2010, CDOs were only very sporadically present in our sample. However, at the end of our observation period, in 2020, a CDO is present in about one-fifth of the analyzed firms. In addition, also the DIGITAL RATIO across the 10-K reports and conference calls has increased strongly over time. In that regard, our data indicates that the average DIGITAL RATIO of the conference calls is significantly higher than the average DIGITAL RATIO of the 10-K reports. Since existing research indicates industrial differences in the frequency of CDO appointments, we also investigate the CDO RATIO and the DIGITAL RATIO per industry. The relevant information is shown in Figure 28. The chart on the left-hand side illustrates the CDO RATIO per industry and its development over the years 2007, 2014, and 2020. The chart on the right-hand side illustrates the average DIGITAL RATIO over the entire study period per industry and communication tool (10-K reports vs. conference calls).

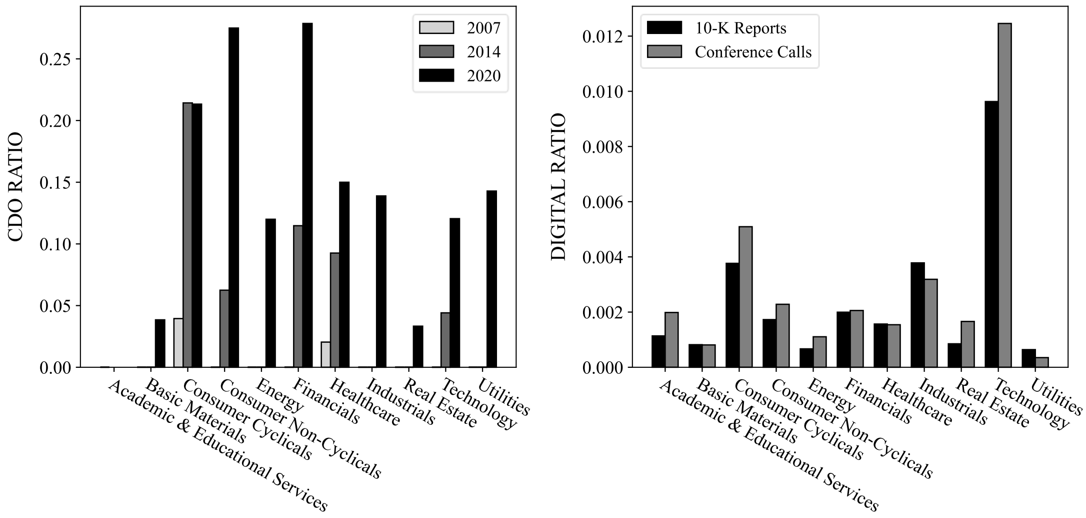

Figure 28. CDO Ratio and Digital Ratio Across Industries and Communication Tools

In line with existing literature (e.g., Firk et al. 2021; Kessel and Graf-Vlachy 2021), our data shows that the proportion of firms with a CDO in their top management team in firms with a high focus on intangible assets (e.g., Financials) is higher than in firms with a high focus on tangible assets (e.g., Basic Materials). The DIGITAL RATIO measured in 10-K reports and conference calls also varies considerably among different industries. It is not surprising that the DIGITAL RATIO of firms within the technology sector is the highest of all industries. This can be justified because these firms focus on developing and selling technology-based products and services. As a result, they have a high technology focus in their reporting. Consumer cyclical firms and industrial firms have a relatively high DIGITAL RATIO as well. Firms of the basic materials industry and the utility industry have the lowest average DIGITAL RATIO. Finally, another interesting finding in this dataset is that the average DIGITAL RATIO in conference calls is higher than in 10-K reports. This might be due to the fact that 10-K reports only allow little flexibility, whereas conference calls also include a larger share of more spontaneous content. Further, since the content conference call documents is not highly regulated, this could indicate that digital transformation is more of a cheap talk for many firms.

The third dataset includes the control variables gathered from Refinitiv Datastream (accounting and price data) and SDC (M&A data). Table 27 shows the descriptive statistics for all variables. We only include a firm-year observation in our analysis if all variables from equation (9) are available (10-K report, conference calls transcripts, and control variables). We further drop singleton observations (firms with only one observation during the observation period) as they do not add within-firm variation to our analysis. This reduces the total number of firm-year observations for the subsequent analysis to 6,456.

	N	Mean	SD	P(0.01)	P(0.99)
DIGITAL RATIO [CC]	6,456	0.0047	0.0100	0.0000	1.0507
DIGITAL RATIO [10-K]	6,456	0.0039	0.0078	0.0000	1.0415
CDO	6,456	0.0649	0.2464	0.0000	1.0000
INTANGIBLES	6,456	0.1667	0.4313	0.0000	1.783
MTB	6,456	2.6742	59.7569	-40.2000	45.9900
TOTAL ASSETS (M)	6,456	41,267	142,956	475.685	731,781
ROA	6,456	0.0697	0.0867	-0.2299	0.2990
LEVERAGE	6,456	0.6177	18.1205	-12.7548	17.1008
RETURN	6,456	0.1545	0.4173	-0.6859	1.4572
DIGITAL M&A	6,456	0.0694	0.3500	0.0000	2.0000
RELATED CxO	6,456	0.7103	0.4536	0.0000	1.0000

Table 27. Descriptive Statistics

To better understand how the utilized variables are interrelated, we calculate the pair-wise correlations (Pearson correlation). The results can be obtained from Table 28.

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(1) DIGITAL RATIO [CC]	1										
(2) DIGITAL RATIO [10-K]	0.68*	1									
(3) CDO	0.14*	0.11*	1								
(4) INTANGIBLES	0.02*	0.04*	0.02*	1							
(5) MTB	0.01	0.02	0.01	0.01	1						
(6) TOTAL ASSETS (M)	-0.01	0.01	0.06*	-0.01	-0.01	1					
(7) ROA	0.10*	0.10*	-0.02	-0.09*	0.01	-0.09*	1				
(8) LEVERAGE	0.01	0.01	0.01	-0.01	0.48*	-0.02*	-0.01	1			
(9) RETURN	0.05*	0.04*	-0.01	-0.01	0.03*	0.04*	0.10*	-0.01	1		
(10) DIGITAL M&A	0.33*	0.29*	0.04*	0.01	0.01	0.03*	0.06*	0.01	0.01	1	
(11) RELATED CxO	0.06*	0.08*	0.08*	-0.01	0.01	0.02*	0.04*	0.01	-0.01	0.07*	1

* significance at the 0.05 level

Table 28. Correlation Matrix

A significant positive correlation between CDO presence and DIGITAL RATIO can be observed, which could indicate a positive relation between CDO presence and the volume of digital transformation-related signals in external corporate communication. The correlation matrix also shows that a higher DIGITAL RATIO is associated with a higher return on assets and higher stock returns. This could be interpreted as communication about digital transformation measures that positively impact profitability (if increased communication is associated with increased digital transformation activities) and investors' assessment. The reversed direction could also be possible, so that particularly profitable firms invest their resources, especially in such activities, and communicate it to the capital market. There is also a positive correlation between firm size and CDO presence which is also as expected because larger firms typically have a larger board (Eisenberg et al. 1998) and are therefore more likely to implement more specific positions such as that of a CDO. Finally, we see higher Digital Ratios for firms that engage in DIGITAL M&A and that have RELATED CxOs in their top management team.

5 Empirical Results

In order to evaluate the impact of CDO presence on the volume of digital transformation-related signals, we make use of the underlying data's panel structure. The analysis is divided into two parts. In the first part, we consider only first-time CDO appointments, while in the second part, the entire set of observations is utilized by the panel regression. First, we only look at the cases of first-time CDO appointments for which we can observe the two years before and the two years after the appointment. For this, the firm must exist for the entire five years, and data on CDO presence, conference call transcripts, and 10-K reports must be available for each year. This results in a total of 81 first-time appointments we can utilize. Thereby, our analysis focuses on the transition from a firm without CDO presence to a firm with CDO presence. The results of this analysis can be obtained from Figure 29.

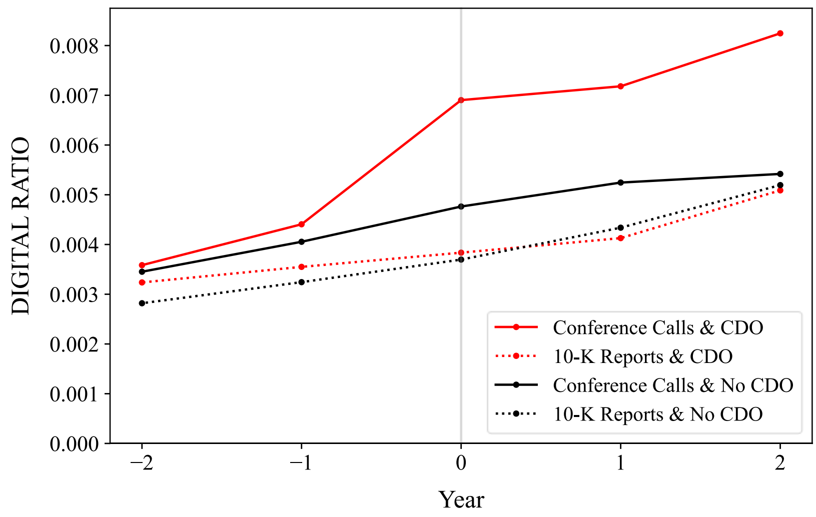

Figure 29. Digital Transformation Activities Around CDO Appointments

The red lines in Figure 29 show how the DIGITAL RATIO changes over time relative to the first-time CDO appointment. The year 0 marks the year in which the CDO is appointed. The black lines serve as references and show how the DIGITAL RATIO developed in the conference calls and the 10-K reports among those firms that did not appoint a CDO during the entire observation period. The temporal structure of the CDO groups and the reference groups are matched. For the conference calls (solid line), we observe an almost identical DIGITAL RATIO in the two years before the appointment. However, in the year of the CDO appointment, the DIGITAL RATIO rises sharply, while the reference firms only follow the overall trend. In the two years after the CDO appointment, the DIGITAL RATIO are again relatively parallel, but those of the firms with a CDO are on a much higher overall level. This suggests that the CDO has triggered an increase in the volume of digital transformation-related signals in conference calls. Based on 10-K reports (dotted line), no such effect can be observed. The two graphs are thus relatively similar over the entire period under consideration. Again, this could indicate that digital transformation is more of a cheap talk for many firms. Firms indeed increase their digital transformation-related signaling after CDO appointments, however, mostly in less regulated communication tools with lower signal reliability. Further, in 10-K reports firms have only little flexibility, whereas conference calls also include a larger share of spontaneous content and Q&A sessions.

The findings derived from Figure 29 provide a first indication that H1 and H2 can be confirmed. Thereby, the confirmation of H1 is mainly driven by the conference calls. To provide statistical evidence, we make use of a research approach in which all 6,456 firm-year combinations are utilized and not only the limited time periods surrounding first-time CDO appointments as in the previous analysis. Due to the larger number of observations and the numerous control variables, this analysis allows more precise statements about the effects of CDOs on digital transformation-related signaling. The results are shown in Table 29.

	[1] DIGITAL RATIO [CC]	[2] DIGITAL RATIO [10-K]	[3] DIGITAL RATIO [CC]	[4] DIGITAL RATIO [10-K]
CDO	0.0053***	0.0035***	0.0031**	0.0009*
	(10.51)	(9.05)	(2.54)	(1.80)
INTANGIBLES			-0.0016**	-0.0011**
			(-2.12)	(-2.19)
MTB			0.000001	0.000001
			(1.15)	(1.51)
ln(TOTAL ASSETS)			0.0014**	0.0013***
			(2.10)	(2.96)
ROA			0.0010	0.00004
			(0.43)	(0.02)
LEVERAGE			-0.000002	-0.000003
			(-0.48)	(-1.08)
RETURN			0.0003	0.0002
			(0.69)	(1.29)
ln(1+DIGITAL M&A)			0.0030***	0.0025**
			(3.22)	(2.43)
RELTATED CxO			-0.0003	-0.0007*
			(-0.32)	(-1.72)
INTERCEPT	0.0043***	0.0037***	-0.0085	-0.0075*
	(33.90)	(37.12)	(-1.43)	(-1.93)
N	6,456	6,456	6,456	6,456
Fixed Effects	No	No	Firm & Year	Firm & Year
Clustering	No	No	Firm	Firm
Adj. R^2	0.017	0.012	0.628	0.718
* p<0.1, ** p<0.05, *** p<0.01; t statistics in parentheses				

Table 29. Panel Regression

Models [1] and [2] are standard OLS regressions. Thus, they do not consider the panel structure of the underlying data. However, they show that CDO presence has a positive effect on the DIGITAL RATIO. Models [3] and [4] correspond to equation (9) specified in the methodology section. Also, based on these regression models, it can be seen that CDO appointments have a significant positive effect on the DIGITAL RATIO in the conference calls (p=0.010). The effect with respect to the 10-K reports is almost significant (p=0.072) but with less than a third of the magnitude compared to the effect on conference calls. Thus, H1 and H2 can be confirmed. Both models account for multiple control variables, firm fixed effects, and year fixed effects. Furthermore, we use heteroskedasticity-robust standard errors that are clustered on the firm dimension. The regression coefficient of 0.0031 for the conference calls also shows an economically significant effect size, considering that the mean value across all years and firms is only 0.0047 and that even if only the firms that introduce a CDO later in the observation period are considered, the mean DIGITAL RATIO is only 0.0055 in the year before the CDO is appointed. The CDO effect thus corresponds to an increase of 56.36% in the DIGITAL RATIO of conference calls. The results clearly confirm that CDO presence is associated with an increase in the volume of digital transformation-related signals and that this signaling primarily takes place via less regulated communication tools (i.e., conference calls). Interestingly, we do not see a significant effect of RELATED CxO positions on digital transformation-related signaling, which emphasizes the specific role of CDOs concerning digital transformation.

6 Discussion

6.1 Theoretical and Practical Implications

This paper enhances existing literature in the research stream "Consequences of CDO presence" in manifold ways. Our analysis shows that CDO presence is continuously increasing across S&P500 firms which underlines the high relevance of CDOs for firms. Further, it underlines the strategic importance of dealing with the decision on appointing a CDO or not. Our data indicates that CDO presence and digital transformation-related signaling in external communication tools vary across industries. In line with existing literature, we show that CDO presence in firms focusing on intangible assets is higher than in firms with tangible assets (e.g., Firk et al. 2021; Kessel and Graf-Vlachy 2021).

Consistent with existing studies, we show that the appointment of a particular position in the top management team can be associated with increased signaling in the area of responsibility of this person (e.g., Kralina 2018). In our case, we show that CDO presence leads to a higher volume of digital transformation-related signals within firms' main external communication tools (i.e., firm's 10-K reports and conference calls). Thus, CDO presence is associated with a higher volume of digital transformation-related signals in a firm's corporate communication tools. In addition, our results indicate significant differences between the volume of digital transformation-related signals in highly regulated communication tools (i.e., 10-K reports) and less regulated ones (i.e., conference calls). Conference calls contain a relatively higher amount of such signals. Concerning signal reliability, it can be assumed that digital transformation is more of a cheap talk for many firms. These firms like to talk about it but do not substantially engage in digital transformation activities. In that regard, CDO presence indeed reinforces digital transformation-related signaling. However, mostly associated with relatively low signal reliability. Overall, it remains questionable if the increased signaling through CDO presence is suitable for reducing potential information asymmetries. Another potential reason for the predominant use of non-regulated communication tools, is that 10-K reports are highly standardized documents in which firms have only little flexibility. This makes it more difficult for firms to address current issues as quickly as possible. Second, conference calls also include a large share of spontaneous content. In such conference calls, firms have the possibility to present and discuss current issues. In addition, in conference calls, external analysts and other persons can ask questions that can increasingly be related to digitalization activities. Therefore, it can be assumed that digital transformation-related signaling works easier through less regulated communication tools. However, again, this bears the risk that such less regulated communication tools are not as trustworthy as more regulated communication tools. Consistent with these results, we also found considerable variation among the different document types concerning the increase of the scope in digital transformation-related content as a direct reaction to the first-time appointment of a CDO. Whereas we can observe a sharp increase in digital transformation-related content in conference calls as a reaction to CDO appointments, the increase in 10-K reports does not exceed the overall trend. This also might be due to the highly regulated and standardized nature of 10-K reports. Overall, these results indicate that less regulated communication tools (i.e., conference calls) are more likely to be used to address digital transformation-related topics. Investors searching for information concerning firms' digital transformation activities, therefore, are more likely to find such information in less-regulated communication tools. However, at the same time, these communication tools are accompanied by lower signal reliability. Thus, it could be that firms rather just referencing digital technologies and digital transformation in order to, say, impress the investors instead of implementing these technologies.

For firms, our study can support the decision-making process when facing the question of appointing a CDO to the top management team or not. Our study suggests that appointing a CDO to the top management team is an excellent option for firms that are at least interested in improving their digital perception with regard to external stakeholders. Nevertheless, our study does not replace a systematic decision-making process. Firms should also consider their specific requirements and determine their individual needs.

Finally, although our results confirm hypothesis 2 that the impact of CDO presence on the quantity of digital transformation-related content in less regulated communication tools is different than in highly regulated communication tools, these results are questionable from a regulatory point of view. On the one hand, firms have to report all material information, including qualitative information, in a 10-K report. Indeed, digital transformation-related topics are material information as the degree of digitalization impacts the future competitiveness of firms. However, on the other hand, digital transformation-related topics play a rather subordinate role in 10-K reports.

6.2 Limitations and Future Research

Besides the careful design of our research approach, this study is subject to some limitations. First, our study only considers S&P500 firms. Therefore, our results can only be generalized to large US-based firms. Future research could build on this by verifying whether our results can be confirmed in other countries and for small and medium-sized enterprises (SMEs). In addition, we only assessed two specific communication tools (i.e., conference calls and 10-K reports). These are the very important communication tools of firms to get in touch with investors and other stakeholders. However, these sources still only represent a selection of relevant communication tools of firms. Future research could adopt this methodology and could, for example, also analyze firms' websites and other publicly available sources.

The volume of digital transformation-related signals in documents is measured based on a dictionary, which allows a high degree of transparency and replicability for future research. However, machine learning techniques may extract such content with a higher degree of accuracy (A.H. Huang et al. 2014a). Further, although we already extended the existing dictionary of digital words, future research could extend it even further, e.g., by adding more digital technology-related word-groups.

One of the most ubiquitous problems in research on firm's management teams concerns endogeneity. Decisions on the structure of the top management team are typically made consciously and in particular based on strategic considerations. As a result, our results may not be causally driven by the CDO. Our results (increased relevance of digital transformation activities) and the appointment of the CDO could also be driven simultaneously by a third variable that is not considered. At the same time, the CDO presence could have no causal impact on the scope of digital transformation-related communication. For this reason, our results can only indicate an association between the presence of a CDO and the relevance of digital transformation activities in firms. While we control for numerous possible factors through the use of firm and year fixed effects as well as control variables that could drive our results, we cannot derive flawless causality based on our study design. This leads to the possibility that the results could be affected by the phenomenon that firms with a higher strategic focus on digital transformation naturally engage more in digital transformation activities (independent of the presence of a CDO). Future research is encouraged to further improve this approach in order to minimize endogeneity issues further.

Our study indicates that CDO presence is associated with a higher volume of digital transformation-related signals quantitatively. Future research could build on this by verifying whether there also is a causal effect between these variables. Further, our study assumes that this higher quantity of digital transformation-related signaling, i.e., higher information quantity, goes along with higher information quality, reducing potential information asymmetries. However, it remains unclear whether the presence of a CDO really has a positive impact on the quality of digital transformation-related signaling and thereby has the power to reduce potential information asymmetries. Future research could build on this by qualitatively analyzing the content of digital transformation-related signaling of firms with a CDO vs. firms without a CDO. Finally, future research could also investigate whether a higher quantity of digital transformation-related signaling has an impact on specific information asymmetry proxies (e.g., bid-ask spreads), financial performance, and capital market parameters.

7 Conclusion

Existing CDO-related research indicates that firms appoint CDOs to the top management team intending to drive and coordinate digital transformation activities and communicate digital transformation-related topics with stakeholders (e.g., R. Grossman and Rich 2012; Singh and Hess 2017). However, until now, it remained unclear whether those firms appointing a CDO are more likely to conduct digital transformation-related signaling, especially in their external communication tools, and whether a CDO appointment is an appropriate instrument to increase a firm’s external visibility and to handle investor relations concerning digital transformation. With this study, we approached this research gap by analyzing the impact of CDO presence on the volume of digital transformation-related signals in firms’ external communication tools.

Our empirical results indicate that CDO presence leads to an increase in the discussion of digital transformation-related topics in firms’ external communication tools. This increase in the volume of digital transformation-related signals can be observed directly after a CDO appointment. In addition, we show that this effect is mainly driven by the impact on less regulated communication tools (i.e., conference calls). Overall, our results highlight that the presence of a CDO in the top management team can be associated with a higher volume of digital transformation-related signals in a firms’ external communication tools. Therefore, it can be concluded that a CDO is an appropriate instrument to increase a firms' external visibility and to handle investor relations concerning digital transformation. However, concerning signal reliability, investors and other external stakeholders need to evaluate whether a firm actually engages in digital transformation or if it is more of a cheap talk.

C. Contributions

1 Summary of Results

In **Research Area I**, the methodology of document similarity is intensively studied. An analysis of the accounting and finance literature revealed that researchers do not follow a uniform approach when calculating document similarity. In Bankamp and Muntermann (2022, paper I.1), the "similarity cube" framework is developed based on literature. The cube comprises three important dimensions of document similarity: temporal, object, and author. This framework is the foundation for the experiment comparing different document representations to answer **Research Question I.1**.

> **Research Question I.1:** How should one choose document representations based on the dimension of similarity to capture?

On the one hand, the document representation "doc2vec" is suited across all dimensions and thus offers a universal option for document representation in similarity measurement. The representation based on the topic model LDA, on the other hand, has a specific field of application and is only suitable for capturing the object dimension. Another important finding is that the simple bag-of-words models achieve satisfying results and are therefore promising choices for projects where replicability and transparency are highly relevant. Another important finding of Bankamp and Muntermann (2022, paper I.1) is the described phenomenon of seasonality, which can be found in document similarity along the temporal dimension.

Research Area II explores information asymmetries on the capital market between companies and investors and how they can be reduced by various means.

> **Research Question II.1:** How should a system be designed for extracting sustainability-relevant information from analyst reports?

To answer this question Bankamp and Muntermann (2021, paper II.1) develop an artifact. The core of this system's design is the combination of a dictionary approach and a machine learning-based BERT model to increase precision and recall and to reduce the problem of class imbalance when training the model.

> **Research Question II.2:** How do analysts modify their reports to respond to the changed market conditions induced by MiFID II?

To reduce information asymmetries, investors can seek the help of financial analysts and gain insights from their reports and assessments of companies in which they are about to invest. The European regulation MiFID II has comprehensively changed the business model of sell-side analysts. This has not only affected the coverage and forecasts of analysts, as other studies have already shown, but also had an impact on the content of the reports, as Bankamp et al. (2022, paper II.2) demonstrate when answering **Research Question II.2.** After the introduction of MiFID II, the reports contain less content that has already been published by peers in their reports. Thus, following MiFID II, analysts have become more independent, engage in less herd behavior, and rely more on their private information than before the regulation. At the same time, the new pricing model minimizes the incentive to write overly optimistic reports. However, other than these positive effects on content, there has also been a decrease in analyst coverage.

> **Research Question II.3**: How does MiFID II's research unbundling impact the provision of information by the media?

In addition to the direct effects of the regulation on analysts, other stakeholders may also have been indirectly affected by the regulation. The results from Bankamp (2022, paper II.3) show that the media coverage on companies decreases if the companies concerned are negatively affected by a MiFID II-induced loss of analyst coverage. In particular, as financial media plays an important role in disseminating information (L. Fang and Peress 2009; Bushee et al. 2010; Drake et al. 2014), the coverage loss penalizes small investors that use these news outlets as an important information medium and do not have access to analysts' reports or direct access to a company's management team. MiFID II's objectives to increase transparency and to protect investors are counteracted, at least for the subgroup of small investors.

> **Research Question II.4a**: How does CDO presence impact the volume of digital transformation-related signals in external communication tools?

In addition to the provision of information by information intermediaries or screening of investors, management may also deliberately share information to reduce asymmetries. According to the signaling theory, companies have an incentive to do so if they have positive characteristics and wish to disclose them to the counterparty. The results in Metzler et al. (2021, paper II.4) on **Research Question II.4a** show that companies share more information about their digital transformation with the capital market after hiring a CDO.

> **Research Question II.4b**: How does the volume of digital transformation-related signals differ across communication tools with different degrees of regulation?

However, the results in Metzler et al. (2021, paper II.4) also indicate that companies especially use less regulated communication channels in the form of conference calls to send these signals (**Research Question II.4b**). In highly regulated 10-K reports, a less steep increase in signal volume can be observed. This result calls into question both signal reliability and whether this form of increased signaling can actually reduce information asymmetries with respect to digital transformation.

2 Implications

Research Area I leads to implications for researchers who want to integrate document similarity as a methodology in their research. While this methodology was not widely used until recently (Loughran and McDonald 2016), the literature review in Bankamp and Muntermann (2022, paper I.1) suggests that it is gaining importance in the finance and accounting context. Researchers can apply the similarity cube developed in Bankamp and Muntermann (2022, paper I.1) to clarify which similarity dimension they want to capture in order to answer their research question. The results from **Research Area I** will also help them to find an appropriate document representation. Researchers should not simply choose the newest and most complex document representations without proper justification. Established and simple bag-of-words models deserve their justification for achieving promising results. Overall, the results also provide a strong foundation for researchers to justify and discuss their choice of representations. For practitioners, the implications described above apply as well.

The results from Bankamp and Muntermann (2021, paper II.1) in **Research Area II** have implications for investors who want to integrate environmental concerns into their investment process and do not want to rely (solely) on sustainability scorings from external providers. This will allow investors to automatically screen text documents for sustainability information and thus qualitatively evaluate the sustainability performance of potential investment objects. In addition to screening during the investment selection process, the system also makes it possible to continuously monitor companies in which the investor has invested and to take countermeasures if necessary. In addition to its practical applications, the system also offers implications for researchers. The system could be used in empirical research to design a measure for the volume of disclosure of environmentally relevant information.

The results from the two studies on MiFID II (Bankamp et al. 2022, paper II.2; Bankamp 2022, paper II.3) have important implications for various stakeholders. Researchers might be interested in the finding of increased novelty, since it could offer an explanation for higher stock market reactions to individual analyst releases, as found in previous studies (e.g., B. Fang et al. 2020). In addition, Bankamp et al. (2022, paper II.2) draw attention to the textual part of analysts' disclosures, which has been ignored by previous work on MiFID II. This provides a starting point for a deeper engagement with the qualitative work of analysts in relation to regulatory change and beyond. The results in Bankamp (2022, paper II.3) emphasize the interconnection between information intermediaries on the capital market and the far-reaching consequences of regulatory interventions. These findings highlight the relevance of researching the secondary effects of regulations.

The results of both studies have important implications for regulators. On the one hand, the European regulator can use these findings to identify the intended and unintended effects of its regulation, compare them with its objectives and, if necessary, adjust the regulation. The fact that the European regulator again allowed research bundling for small and medium-sized enterprises in February 2021 in response to the pandemic (Directive (EU) 2021/338; EU (2021)) shows that it is reviewing and adapting its regulation on an ongoing basis. On the other hand, the results are also interesting for non-European regulators who are considering similar adjustments to their legislation. The coverage loss documented in Bankamp (2022, paper II.3) is particularly relevant from a regulatory point of view, as private investors with no access to other sources depend on the one hand on information provided by financial journalists and on the other hand on the protection from regulators that should provide a level playing field to all market participants. The results in Bankamp et al. (2022, paper II.2) might also help investment practitioners who are responsible for sourcing analyst research. As the novelty of the reports has increased after MiFID II implementation, relying on sell-side research might be more rewarding than prior to MiFID II. In this context, it could also be beneficial to analyze reports from different analysts, as the overlaps have become smaller.

Furthermore, the composition of the top management team is an important strategic decision that companies face. This decision also includes the question of whether a CDO should be appointed. The results in Metzler et al. (2021, paper II.4) can support companies here. On the one hand, they show that appointing a CDO can increase the external communication of the company’s digital transformation. On the other hand, they demonstrate that this information only increases slightly in highly regulated financial reports. Since the communication of digital transformation is primarily conducted via less regulated documents, it raises the question of whether the words are followed by actions in the digital transformation. Companies should ensure that the digital transformation processes are actually implemented by or with the CDO.

3 Limitations

Research Area I has its limitations. Even though this area is limited to finance and accounting literature, the experiment cannot cover these research fields completely. In the experiment, analyst reports are used as a data basis. However, in the field of finance and accounting research, many other texts, such as financial reports, conference calls, and press releases, are of interest. It cannot be ruled out that the results may differ when other documents are used. Furthermore, the paper examines only one (albeit important) step in the calculation of document similarity, namely document representation. Other important steps that are not considered here are the choice of pre-processing and the choice of the distance measure. In addition, only a selection of representations was used for the experiment. Due to the abundance of available representations, some may achieve even better results than those used in this study. Finally, the similarity cube is not complete. Other dimensions as well as subdimensions are conceivable. Style, for example, could be a subdimension of the author dimension.

A general limitation of **Research Area II** is that its results cannot prove that the examined systems (Bankamp and Muntermann 2021, paper II.1) and effects (remaining papers from Research Area II) that could influence information asymmetries, actually alter information asymmetries between investors and management. The results in Bankamp and Muntermann (2021, paper II.1) cannot quantify the usefulness of the extracted information for decision-making. In addition, the scope of the developed prototype is relatively narrow, as only the ecological dimension of sustainability is used, while the economic and social dimensions are not considered. Furthermore, only analyst reports are used as representatives of financial documents. The design principles' generalizability with respect to other sustainability dimensions and documents must be proven first before applying them to a broader set of problems.

The two studies on MiFID II (Bankamp et al. 2022, paper II.2; Bankamp 2022, paper II.3) are subject to some general assumptions and limitations. Because the research unbundling of MiFID II does not depend on the covered company's jurisdiction and MiFID II was found to have a global impact (Allen 2019), the actual effect of MiFID II will affect not only the treatment group but also to a certain extent the control group in Bankamp et al. (2022, paper II.2) and Bankamp (2022, paper II.3). The problem of separating the pre- and post-MIFID II periods also exists. Although the implementation date of the regulation was clearly defined, market participants were informed about this long before. Analysts and journalists were therefore able to adjust their behavior before the regulation came into effect and anticipate the regulatory change. In Bankamp et al. (2022, paper II.2), analyst's optimism is measured using sentiment polarity. The weakness of this method lies in the problem of a missing calibration point of this text-based measure. The analysis allows to make conclusions about analysts becoming less optimistic in their communication due to MiFID II, but not that they were generally too optimistic before. This is different to the assessment of optimism in price targets, were the forecast can be compared to the realized price value that is later observed. For sentiment there is no way of comparing the analyst's textual assessment (sentiment) with the value that actually occurred. The two studies are also subject to limitations in terms of data sources. In Bankamp et al. (2022, paper II.2), only a relatively small number of companies (i.e., 88) were examined. Due to their size, the companies have comprehensive analyst coverage, which is necessary to analyze the similarity between reports from different analysts. At the same time, the generalizability of the results (especially with regard to smaller companies) may be limited. The analyst report database in Thomson ONE is also neither complete nor randomly selected. As analysts explicitly decide whether and what they provide via the database (Amiram et al. 2018), it could reduce the data's representativeness. Moreover, the I/B/E/S database containing analyst forecasts and the RavenPack database for news media (both utilized in Bankamp (2022, paper II.3)) do not cover all analysts or media outlets, which could again reduce generalizability.

Endogeneity is a fundamental problem in research evaluating the effects of the top management team, as in Metzler et al. (2021, paper II.4). The decision to appoint a certain board position is usually a conscious decision of the company. It is therefore possible that both the decision to appoint a CDO and the decision to engage more in digital transformation-related signaling are driven by a third variable. It was tried to minimize this problem by using numerous control variables and fixed effects, but it cannot be made a definitive statement about a causal relationship. In addition, there are limitations in Metzler et al. (2021, paper II.4) that restrict the generalizability of the results. The S&P 500 was selected as the sample, so the results may not be applicable to small companies or companies outside the US. In particular, the 10-K reports are very standardized; therefore, the results on signaling in the financial reports could be different when analyzing jurisdictions where the financial reports are less standardized. The two document types (financial reports and conference calls) also do not represent the entire capital market-related communication. Thus, signaling could also be carried out via other channels, such as press releases, which were not included in the study. The dictionary-based approach was used to construct the measure of digital transformation-related signaling. Machine learning-based methods might provide higher accuracy (A.H. Huang et al. 2014a) but are harder to replicate.

4 Future Research

Research Area I offers numerous starting points for subsequent research. The similarity cube offers an excellent opportunity for extension of further dimensions or sub-dimensions and thus to obtain a better and more refined understanding of document similarity. Further research opportunities lie in the investigation for other document classes. The analysis of financial reports might be a promising starting point here. Furthermore, the analysis of additional document representations or the enrichment of the representation with additional information, for example named-entity recognition as in Friburger et al. (2002), could be a fruitful approach. The scope of subsequent analysis could also be extended by including aspects of pre-processing or distance measures and showing the dependencies of the individual pre-processing steps. In addition to this methodological research path, the research area should also motivate new empirical work that uses document similarity as a method to answer many interesting research questions.

As mentioned in the limitations (Section 3), **Research Area II** overall could be extended by examining whether and to what extent the investigated system and effects actually impact information asymmetries. This could be done by empirically testing market-based indicators such as liquidity (Lev 1988). For the system proposed in Bankamp and Muntermann (2021, paper II.1), a field study with (institutional) investors could be a promising extension of the current evaluation. In this way, the usefulness of the system for investors' decision-making could be tested in a real-world scenario. Further extensions for future research lie in the adaptation and application of the design principles to different problems and domains. The problem of information extraction is a fundamental one and is exacerbated by the increasing amount of unstructured (text) data. The extent to which the proposed design principles are applicable here is an important question for future research. In addition to these cross-domain extensions, future research could also develop the functionalities of the proposed system.

The introduction of MiFID II has triggered many studies on the reaction of sell-side analysts. The two papers on MiFID II in this thesis may provide an impetus for somewhat broader research that also examines the qualitative work of analysts and explores the effects on other information intermediaries. Prior research on MiFID II has focused almost exclusively on quantitative analyst output. Bankamp et al. (2022, paper II.2) may stimulate regulatory-related research that more closely examines the effects on analysts' qualitative opinion. Interesting questions here could concern, for example, the readability (e.g., De Franco et al. 2015) of reports or their topic composition (e.g., A.H. Huang et al. 2018). The results from Bankamp (2022, paper II.3) serve as a starting point for research examining secondary effects on investors or other information intermediaries. Another interesting question is the extent to which the content of news articles has changed because of the introduction of MiFID II. The methods applied in Bankamp et al. (2022, paper II.2) could also be applied to the textual content of news articles.

The consequences of CDO presence as well as the impact on topic-specific signaling leave much room for subsequent research. First, the problem of endogeneity should be addressed. This could be done by relying on an appropriate natural experiment. In addition, the analysis should be extended along different dimensions to increase generalizability. On the one hand, companies of different sizes and jurisdictions could be included. On the other hand, it would also be interesting to include other documents (such as press releases) that the company might use for signaling purposes. Finally, a qualitative analysis of digital-transformation related signals could accompany the quantitative investigation on signaling volume to better understand the exact information that is brought to the capital market.

References

Adosoglou, G., Lombardo, G., and Pardalos, P. M. 2021. “Neural network embeddings on corporate annual filings for portfolio selection.” *Expert Systems with Applications* 164: 114053.

Aerts, W., Cormier, D., and Magnan, M. 2008. “Corporate environmental disclosure, financial markets and the media: An international perspective.” *Ecological Economics* 64 (3): 643-659.

Agarwal, S., Chen, V. Y. S., and Zhang, W. 2016. “The Information Value of Credit Rating Action Reports: A Textual Analysis.” *Management Science* 62 (8): 2218-2240.

Agrawal, A., Chadha, S., and Chen, Mark A. 2006. “Who Is Afraid of Reg FD? The Behavior and Performance of Sell-Side Analysts Following the SEC’s Fair Disclosure Rules.” *The Journal of Business* 79 (6): 2811-2834.

Aguilar, J. 2018. *We think 3M is a wide-moat business but also think the market has recognized its earning power.* Morningstar Equity Research.

Akerlof, G. A. 1970. “The Market for "Lemons": Quality Uncertainty and the Market Mechanism.” *The Quarterly Journal of Economics* 84 (3): 488-500.

Albaugh, Q., Soroka, S., Joly, J., Loewen, P., Sevenans, J., and Walgrave, S. 2014. “Comparing and Combining Machine Learning and Dictionary-Based Approaches to Topic Coding.” *7th Annual Comparative Agendas Project (CAP) Conference*. Konstanz, Germany.

Allen, J. C. 2019. “US Payment for Research and MiFID II 2019 Survey.” CFA Institute. Accessed 03-29-2022. https://www.cfainstitute.org/en/research/survey-reports/us-payment-research-mifid-ii-2019-survey.

Allen, J. C., and Gellasch, T. 2019. “The Future of Research in the US after MiFID II.” CFA Institute. Accessed 03-29-2022. https://www.cfainstitute.org/en/advocacy/policy-positions/future-research-us-after-mifid-ii.

Amel-Zadeh, A., and Serafeim, G. 2018. “Why and How Investors Use ESG Information: Evidence from a Global Survey.” *Financial Analysts Journal* 74 (3): 87-103.

Amiram, D., Bozanic, Z., Bradshaw, M., and Rozenbaum, O. 2018. “Selective Information Channels for Investment Research.” Working Paper.

Amzallag, A., Guagliano, C., and Lo Passo, V. 2021. “MiFID II Research Unbundling – impact on EU equity markets.” Working Paper. ESMA Working Paper.

Anadu, K., Kruttli, M., McCabe, P., and Osambela, E. 2020. “The Shift from Active to Passive Investing: Risks to Financial Stability?” *Financial Analysts Journal* 76 (4): 23-39.

Angrist, J. D., and Pischke, J.-S. 2008. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton, United States: Princeton University Press.

Anselmi, G., and Petrella, G. 2021. “Regulation and stock market quality: The impact of MiFID II provision on research unbundling.” *International Review of Financial Analysis* 76: 101788.

Austin, P. C. 2011. “Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies.” *Pharmaceutical Statistics* 10 (2): 150-161.

Bailey, W., Li, H., Mao, C. X., and Zhong, R. 2003. “Regulation Fair Disclosure and Earnings Information: Market, Analyst, and Corporate Responses.” *The Journal of Finance* 58 (6): 2487-2514.

Bankamp, S. 2022. “MiFID II and its Unintended Effects on Media Coverage.” Working Paper. Prior version accepted and presented at the 12th Annual Pre-ICIS Workshop on Accounting Information Systems, Austin, United States, 2021 and at the 44th Annual Congress of the European Accounting Association, Bergen, Norway, 2022.

Bankamp, S., and Muntermann, J. 2021. “Are My Stocks Sustainable? Design Principles for Leveraging Information from Analyst Reports.” *Proceedings of the Pacific Asia Conference on Information Systems*. Dubai, United Arab Emirates.

Bankamp, S., and Muntermann, J. 2022. “Understanding the Role of Document Representations in Similarity Measurement in Finance and Accounting.” *Proceedings of the Pacific Asia Conference on Information Systems*. Taipei/Sydney, Taiwan/Australia.

Bankamp, S., Palmer, M., and Muntermann, J. 2022. “MiFID II and its Effects on Analyst Reports.” Working Paper. Prior version accepted and presented at INQUIRE Autumn Residential Seminar, Bath, United Kingdom, 2021.

Bär, D., Zesch, T., and Gurevych, I. 2011. “A Reflective View on Text Similarity.” *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Hissar, Bulgaria.

Barber, B. M., Lehavy, R., McNichols, M., and Trueman, B. 2006. “Buys, holds, and sells: The distribution of investment banks’ stock ratings and the implications for the profitability of analysts’ recommendations.” *Journal of Accounting and Economics* 41 (1): 87-117.

Baron, D. P., and Besanko, D. 1987. “Monitoring, Moral Hazard, Asymmetric Information, and Risk Sharing in Procurement Contracting.” *The RAND Journal of Economics* 18 (4): 509-532.

Beaupain, R., and Girard, A. 2020. “The value of understanding central bank communication.” *Economic Modelling* 85: 154-165.

Bender, M., Clapham, B., Gomber, P., and Koch, J.-A. 2021. “To Bundle or Not to Bundle? A Review of Soft Commissions and Research Unbundling.” *Financial Analysts Journal* 77 (3): 69-92.

Benlian, A., and Hess, T. 2011. “The Signaling Role of IT Features in Influencing Trust and Participation in Online Communities.” *International Journal of Electronic Commerce* 15 (4): 7-56.

Berg, F., Kölbel, J. F., and Rigobon, R. 2020. “Aggregate Confusion: The Divergence of ESG Ratings.” Working Paper. Available at SSRN 3438533.

Berman, S., Baird, C. H., Eagan, K., and Marshall, A. 2020. “What makes a Chief Digital Officer successful?” *Strategy & Leadership* 48 (2): 32-38.

Beyer, A., and Guttman, I. 2011. “The Effect of Trading Volume on Analysts’ Forecast Bias.” *The Accounting Review* 86 (2): 451-481.

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., and Venkatraman, N. 2013. “Digital business strategy : toward a next generation of insights.” *MIS Quarterly* 37 (2): 471-482.

Bhargava, H. K., and Choudhary, V. 2004. “Economics of an Information Intermediary with Aggregation Benefits.” *Information Systems Research* 15 (1): 22-36.

Bhattacharya, S. 1979. “Imperfect Information, Dividend Policy, and "The Bird in the Hand" Fallacy.” *The Bell Journal of Economics* 10 (1): 259-270.

Björkdahl, J. 2020. “Strategies for Digitalization in Manufacturing Firms.” *California Management Review* 62 (4): 17-36.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3: 993-1022.

Boulding, W., and Kirmani, A. 1993. “A Consumer-Side Experimental Examination of Signaling Theory: Do Consumers Perceive Warranties as Signals of Quality?” *Journal of Consumer Research* 20 (1): 111-123.

Bowman, E. H. 1984. “Content Analysis of Annual Reports for Corporate Strategy and Risk.” *Interfaces* 14 (1): 61-71.

Bradshaw, M. T., Ertimur, Y., and O'Brien, P. 2017. “Financial Analysts and Their Contribution to Well-Functioning Capital Markets.” *Foundations and Trends® in Accounting* 11 (3): 119-191.

Bradshaw, M. T., Huang, A. G., and Tan, H. 2019. “The Effects of Analyst-Country Institutions on Biased Research: Evidence from Target Prices.” *Journal of Accounting Research* 57 (1): 85-120.

Brav, A., Graham, J. R., Harvey, C. R., and Michaely, R. 2005. “Payout policy in the 21st century.” *Journal of Financial Economics* 77 (3): 483-527.

Brav, A., and Lehavy, R. 2003. “An Empirical Analysis of Analysts' Target Prices: Short-term Informativeness and Long-term Dynamics.” *The Journal of Finance* 58 (5): 1933-1967.

Brown, S. V., Hillegeist, S. A., and Lo, K. 2004. “Conference calls and information asymmetry.” *Journal of Accounting and Economics* 37 (3): 343-366.

Brown, S. V., and Knechel, W. R. 2016. “Auditor-Client Compatibility and Audit Firm Selection.” *Journal of Accounting Research* 54 (3): 725-775.

Brown, S. V., and Tucker, J. W. 2011. “Large-Sample Evidence on Firms’ Year-over-Year MD&A Modifications.” *Journal of Accounting Research* 49 (2): 309-346.

Bushee, B. J., Core, J. E., Guay, W., and Hamm, S. J. W. 2010. “The Role of the Business Press as an Information Intermediary.” *Journal of Accounting Research* 48 (1): 1-19.

Call, A. C., Emett, S. A., Maksymov, E., and Sharp, N. Y. 2022. “Meet the press: Survey evidence on financial journalists as information intermediaries.” *Journal of Accounting and Economics* 73: 101455.

Cameron, A. C., and Miller, D. L. 2015. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources* 50 (2): 317-372.

Campbell, T. S., and Kracaw, W. A. 1980. “Information Production, Market Signalling, and the Theory of Financial Intermediation.” *The Journal of Finance* 35 (4): 863-882.

Cannon, J. N., Ling, Z., Wang, Q., and Watanabe, O. V. 2020. “10-K Disclosure of Corporate Social Responsibility and Firms’ Competitive Advantages.” *European Accounting Review* 29 (1): 85-113.

Card, D., and Krueger, A. B. 1994. “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania.” *The American Economic Review* 84 (4): 772-793.

Cavallo, E., Powell, A., and Rigobon, R. 2013. “Do Credit Rating Agencies Add Value? Evidence from the Sovereign Rating Business ” *International Journal of Finance & Economics* 18 (3): 240-265.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., and Tar, C. 2018a. “Universal Sentence Encoder for English.” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Céspedes, M., Yuan, S., and Tar, C. 2018b. "Universal Sentence Encoder." Working Paper. Available at arXiv:1803.11175.

Cerin, P. 2010. "Analysing the Environmental Content of Financial Analyst Reports by developing an ESG Framework that incorporates Business Opportunities and the Product Perspectives." Working Paper. Sustainable Investment and Corporate Governance Working Papers.

Certo, S. T. 2003. "Influencing Initial Public Offering Investors with Prestige: Signaling with Board Structures." *The Academy of Management Review* 28 (3): 432-446.

Chauvey, J.-N., Giordano-Spring, S., Cho, C. H., and Patten, D. M. 2015. "The Normativity and Legitimacy of CSR Disclosure: Evidence from France." *Journal of Business Ethics* 130 (4): 789-803.

Chen, C. Y., and Chen, P. F. 2009. "NASD Rule 2711 and Changes in Analysts' Independence in Making Stock Recommendations." *The Accounting Review* 84 (4): 1041-1071.

Chen, J., and Sarkar, S. 2020. "A Semantic Approach to Financial Fundamentals." *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*. Kyoto, Japan.

Chen, W., and Srinivasan, S. 2019. "Going Digital: Implications for Firm Value and Performance." Working Paper.

Chen, Y. 1997. "Equilibrium Product Bundling." *The Journal of Business* 70 (1): 85-103.

Chou, C.-C., Chang, C. J., Chin, C.-L., and Chiang, W.-T. 2018. "Measuring the Consistency of Quantitative and Qualitative Information in Financial Reports: A Design Science Approach." *Journal of Emerging Technologies in Accounting* 15 (2): 93-109.

Cohen, A. 2005. "Asymmetric Information and Learning: Evidence from the Automobile Insurance Market." *The Review of Economics and Statistics* 87 (2): 197-207.

Cohen, A., and Siegelman, P. 2010. "Testing for Adverse Selection in Insurance Markets." *Journal of Risk and Insurance* 77 (1): 39-84.

Cohen, B. D., and Dean, T. J. 2005. "Information Asymmetry and Investor Valuation of IPOs: Top Management Team Legitimacy as a Capital Market Signal." *Strategic Management Journal* 26 (7): 683-690.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, United States: Lawrence Erlbaum Associates.

Cohen, L., Malloy, C., and Nguyen, Q. 2020. "Lazy Prices." *The Journal of Finance* 75 (3): 1371-1415.

Connelly, B. L., Certo, S. T., Ireland, R. D., and Reutzel, C. R. 2011. "Signaling Theory: A Review and Assessment." *Journal of Management* 37 (1): 39-67.

Coval, J. D., and Moskowitz, T. J. 1999. "Home Bias at Home: Local Equity Preference in Domestic Portfolios." *The Journal of Finance* 54 (6): 2045-2073.

Cowen, A., Groysberg, B., and Healy, P. 2006. "Which types of analyst firms are more optimistic?" *Journal of Accounting and Economics* 41 (1): 119-146.

Davila, A., Foster, G., and Gupta, M. 2003. "Venture capital financing and the growth of startup firms." *Journal of Business Venturing* 18 (6): 689-708.

De Amicis, C., Falconieri, S., and Tastan, M. 2021. "Sentiment analysis and gender differences in earnings conference calls." *Journal of Corporate Finance* 71: 101809.

De Franco, G., Hope, O. K., Vyas, D., and Zhou, Y. 2015. "Analyst Report Readability." *Contemporary Accounting Research* 32 (1): 76–104.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41 (6): 391-407.

Deng, Q., Hine, M., Ji, S., and Sur, S. 2017. "Building an environmental sustainability dictionary for the IT industry." *Proceedings of the 50th Hawaii International Conference on System Sciences*. Waikoloa Village, United States.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of NAACL-HLT*. Minneapolis, United States.

Diether, K. B., Malloy, C. J., and Scherbina, A. 2002. "Differences of Opinion and the Cross Section of Stock Returns." *The Journal of Finance* 57 (5): 2113-2141.

Dimson, E., Marsh, P., and Staunton, M. 2020. "Divergent ESG Ratings." *The Journal of Portfolio Management* 47 (1): 75-87.

Dong, W., Liao, S., and Liang, L. 2016. "Financial statement fraud detection using text mining: A systemic functional linguistics theory perspective." *Proceedings of the Pacific Asia Conference on Information Systems*. Chiayi, Taiwan.

Drake, M. S., Guest, N. M., and Twedt, B. J. 2014. "The Media and Mispricing: The Role of the Business Press in the Pricing of Accounting Information." *The Accounting Review* 89 (5): 1673-1701.

Drechsler, K., Wagner, H.-T., and Reibenspiess, V. A. 2019. "Risk and Return of Chief Digital Officers' Appointment – An Event Study." *Proceedings of the International Conference on Information Systems*. Munich, Germany.

Dugar, A., and Nathan, S. 1995. "The Effect of Investment Banking Relationships on Financial Analysts' Earnings Forecasts and Investment Recommendations." *Contemporary Accounting Research* 12 (1): 131-160.

Dyner, I., and Franco, C. J. 2004. "Consumers' Bounded Rationality: The Case of Competitive Energy Markets." *Systems Research and Behavioral Science* 21 (4): 373-389.

Eckardt, M. 2007. *Insurance Intermediation: An Economic Analysis of the Information Services Market*. Heidelberg, Germany: Physica-Verlag.

Eden, R., Burton-Jones, A., Casey, V., and Draheim, M. 2019. "Digital Transformation Requires Workforce Transformation." *MIS Quarterly Executive* 18 (1): 1-17.

Egami, N., and Yamauchi, S. 2021. "Using Multiple Pre-treatment Periods to Improve Difference-in-Differences and Staggered Adoption Design." Working Paper. Available at arXiv: 2102.09948.

Eickhoff, M. 2015. "Enabling Reproducible Sentiment Analysis: A Hybrid Domain-Portable Framework for Sentiment Classification." In *New Horizons in Design Science: Broadening the Research Agenda. DESRIST 2015*, edited by B. Donnellan, M. Helfert, J. Kenneally, D. VanderMeer, M. Rothenberger and R. Winter, 215-229. Cham, Switzerland: Springer International Publishin.

Eickhoff, M., and Muntermann, J. 2016. "How to Conquer Information Overload? Supporting Financial Decisions by Identifying Relevant Conference Call Topics." *Proceedings of the 20th Pacific Asia Conference On Information Systems*. Chiayi, China.

Eisenberg, T., Sundgren, S., and Wells, M. T. 1998. "Larger board size and decreasing firm value in small firms." *Journal of Financial Economics* 48 (1): 35-54.

Ellinghorst, A., Yoward, M., and McNally, C. 2019. *Time for the Shareholders.* Evercore ISI.

Escrig-Olmedo, E., Fernández-Izquierdo, M. Á., Ferrero-Ferrero, I., Rivera-Lirio, J. M., and Muñoz-Torres, M. J. 2019. "Rating the Raters: Evaluating how ESG Rating Agencies Integrate Sustainability Principles." *Sustainability* 11 (3): 915.

Espahbodi, H., Espahbodi, P., and Espahbodi, R. 2015. "Did Analyst Forecast Accuracy and Dispersion Improve after 2002 Following the Increase in Regulation?" *Financial Analysts Journal* 71 (5): 20-37.

EU. 2004. "Directive 2004/39/EC of the European Parliament and of the Council of 21 April 2004 on markets in financial instruments." *Official Journal of the European Union* L 145/1.

EU. 2014. "Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments." *Official Journal of the European Union* L 173/349.

EU. 2021. "Directive (EU) 2021/338 of the European Parliament and of the Council of 16 February 2021." *Official Journal of the European Union* L 68/14.

Fama, E. F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance* 25 (2): 383-417.

Fang, B., Hope, O.-K., Huang, Z., and Moldovan, R. 2020. "The effects of MiFID II on sell-side analysts, buy-side analysts, and firms." *Review of Accounting Studies* 25 (3): 855-902.

Fang, L., and Peress, J. 2009. "Media Coverage and the Cross-section of Stock Returns." *The Journal of Finance* 64 (5): 2023-2052.

Farrell, M., Green, T. C., Russell, J., and Markov, S. 2018. "The Democratization of Investment Research: Implications for Retail Investor Profitability and Firm Liquidity." Working Paper. Available at SSRN: 3222841.

Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. 2010. "Management's tone change, post earnings announcement drift and accruals." *Review of Accounting Studies* 15 (4): 915-953.

Fellbaum, C. 1998. *WordNet: An electronic lexical database*. Cambridge, United States: MIT press.

Feng, L., and Seasholes, M. S. 2004. "Correlated Trading and Location." *The Journal of Finance* 59 (5): 2117-2144.

Ferry, E., and Letzeler, J. 2018. *ABInBev CMD: more brains than brawn.* Exane BNP Paribas Research.

Firk, S., Hanelt, A., Oehmichen, J., and Wolff, M. 2021. "Chief Digital Officers: An Analysis of the Presence of a Centralized Digital Transformation Role." *Journal of Management Studies* 58 (7): 1800-1831.

Fisher, I. E., Garnsey, M. R., and Hughes, M. E. 2016. "Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research." *Intelligent Systems in Accounting, Finance and Management* 23 (3): 157-214.

Flammer, C. 2013. "Corporate Social Responsibility and Shareholder Reaction: The Environmental Awareness of Investors." *Academy of Management Journal* 56 (3): 758-781.

Francis, J. 1997. "Discussion of Self-Selection and Analyst Coverage." *Journal of Accounting Research* 35: 201-208.

Friburger, N., Maurel, D., and Giacometti, A. 2002. "Textual similarity based on proper names." *Proceedings of Workshop on Mathematical Formal Methods in Information Retrieval at th 25th ACM SIGIR Conference*. Tampere, Finland.

Friede, G., Busch, T., and Bassen, A. 2015. "ESG and financial performance: aggregated evidence from more than 2000 empirical studies." *Journal of Sustainable Finance & Investment* 5 (4): 210-233.

Fröberg, E., and Halling, M. 2021. "The Unbundling of Mutual Funds' Trading and Research Commissions: Have Investors Benefited?" Working Paper. Available at SSRN 3892441.

Fu, R., Kraft, A., and Zhang, H. 2012. "Financial reporting frequency, information asymmetry, and the cost of equity." *Journal of Accounting and Economics* 54 (2): 132-149.

Gardiner, R. 2017. *Wendel places part of its stake, creates another overhang.* Davy Research.

Global Sustainable Investment Alliance. 2018. "Global Sustainable Investment Review." Accessed 03-29-2022. http://www.gsi-alliance.org/wp-content/uploads/2019/06/GSIR_Review2018F.pdf.

González, J. R. C., Romero, J. J. F., Guerrero, M. G., and Calderón, F. 2015. "Multi-class multi-tag classifier system for StackOverflow questions." *2015 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*. Ixtapa, Mexico.

Goudjil, M., Koudil, M., Bedda, M., and Ghoggali, N. 2018. "A Novel Active Learning Method Using SVM for Text Classification." *International Journal of Automation and Computing* 15 (3): 290-298.

Gregor, S., and Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact." *MIS Quarterly* 37 (2): 337-355.

Grennan, J., and Michaely, R. 2020. "Artificial Intelligence and High-Skilled Work: Evidence from Analysts." Working Paper. Available at SSRN: 3681574.

Grennan, J., and Michaely, R. 2021. "FinTechs and the Market for Financial Analysis." *Journal of Financial and Quantitative Analysis* 56 (6): 1877-1907.

GRI. 2020. "Consolidated Set of GRI Sustainability Reporting Standards 2020." Accessed 03-29-2022. https://www.globalreporting.org/how-to-use-the-gri-standards/gri-standards-english-language/.

Grossman, R., and Rich, J. 2012. "The Rise of the Chief Digital Officer." Accessed 03-29-2022. https://www.russellreynolds.com/en/insights/reports-surveys/the-rise-of-the-chief-digital-officer.

Grossman, S. J., and Hart, O. D. 1983. "An Analysis of the Principal-Agent Problem." *Econometrica* 51 (1): 7-45.

Guest, N. M., and Kim, J. 2022. "The Media Response to Deterioration in a Firm's Information Environment." Working Paper. Available at SSRN: 3337116.

Gull, R., Shoaib, U., Rasheed, S., Abid, W., and Zahoor, B. 2016. "Pre Processing of Twitter's Data for Opinion Mining in Political Context." *Procedia Computer Science* 96: 1560-1570.

Gunasekarage, A., and Power, D. M. 2002. "The post–announcement performance of dividend–changing companies: The dividend–signalling hypothesis revisited." *Accounting & Finance* 42 (2): 131-151.

Guo, Y., and Mota, L. 2021. "Should Information be Sold Separately? Evidence from MiFID II." *Journal of Financial Economics* 142 (1): 97-126.

Gupta, A., Dengre, V., Kheruwala, H. A., and Shah, M. 2020. "Comprehensive review of text-mining applications in finance." *Financial Innovation* 6 (1): 39.

Haffke, I., Kalgovas, B. J., and Benlian, A. 2016. "The Role of the CIO and the CDO in an Organization's Digital Transformation." *Proceedings of the International Conference on Information Systems*. Dublin, Ireland.

Hanelt, A., Firk, S., Hildebrandt, B., and Kolbe, L. M. 2021. “Digital M&A, digital innovation, and firm performance: an empirical investigation.” *European Journal of Information Systems* 30 (1): 3-26.

Hanley, K. W., and Hoberg, G. 2010. “The Information Content of IPO Prospectuses.” *The Review of Financial Studies* 23 (7): 2821-2864.

Hartzmark, S. M., and Sussman, A. B. 2019. “Do Investors Value Sustainability? A Natural Experiment Examining Ranking and Fund Flows.” *The Journal of Finance* 74 (6): 2789-2837.

Hayes, R. M. 1998. “The Impact of Trading Commission Incentives on Analysts' Stock Coverage Decisions and Earnings Forecasts.” *Journal of Accounting Research* 36 (2): 299-320.

Healy, P. M., and Palepu, K. G. 2001. “Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature.” *Journal of Accounting and Economics* 31 (1): 405-440.

Henry, E. 2008. “Are Investors Influenced By How Earnings Press Releases Are Written?” *The Journal of Business Communication* 45 (4): 363-407.

Herrmann, D. R., Hope, O.-K., and Thomas, W. B. 2008. “International Diversification and Forecast Optimism: The Effects of Reg FD.” *Accounting Horizons* 22 (2): 179-197.

Hess, T., Matt, C., Benlian, A., and Wiesböck, F. 2016. “Options for Formulating a Digital Transformation Strategy.” *MIS Quarterly Executive* 15 (2): 123-139.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. “Design Science in Information Systems Research.” *MIS Quarterly* 28 (1): 75-105.

Heywood, J. S., and Wei, X. 2004. “Education and Signaling: Evidence from a Highly Competitive Labor Market.” *Education Economics* 12 (1): 1-16.

Hilgert, R. 2015. *Volkswagen Update on Diesel Scandal Provides Little New Information; EUR 190 Fair Value Unchanged.* Morningstar Equity Research.

Ho, K.-Y., Shi, Y., and Zhang, Z. 2013. “How does news sentiment impact asset volatility? Evidence from long memory and regime-switching approaches.” *The North American Journal of Economics and Finance* 26: 436-456.

Ho, S. Y., Choi, K. W., and Yang, F. 2019. “Harnessing Aspect-Based Sentiment Analysis: How Are Tweets Associated with Forecast Accuracy?” *Journal of the Association for Information Systems* 20 (8): 1174-1209.

Hoberg, G., and Phillips, G. 2016. “Text-Based Network Industries and Endogenous Product Differentiation.” *Journal of Political Economy* 124 (5): 1423-1465.

Hong, H., Kubik, J. D., and Solomon, A. 2000. “Security Analysts’ Career Concerns and Herding of Earnings Forecasts.” *The RAND Journal of Economics* 31 (1): 121-144.

Hope, O.-K. 2003. “Disclosure Practices, Enforcement of Accounting Standards, and Analysts' Forecast Accuracy: An International Study.” *Journal of Accounting Research* 41 (2): 235-272.

Huang, A. 2008. “Similarity measures for text document clustering.” *Proceedings of the sixth New Zealand Computer Science Research Student Conference*. Christchurch, New Zealand.

Huang, A. H., Lehavy, R., Zang, A. Y., and Zheng, R. 2018. “Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach.” *Management Science* 64 (6): 2833-2855.

Huang, A. H., Zang, A. Y., and Zheng, R. 2014a. “Evidence on the Information Content of Text in Analyst Reports.” *The Accounting Review* 89 (6): 2151-2180.

Huang, X., Teoh, S. H., and Zhang, Y. 2014b. “Tone Management.” *The Accounting Review* 89 (3): 1083-1113.

IIRC & Kirchhoff. 2020. “Closing the gap: The role of integrated reporting in communicating a company’s value creation to investors.” Accessed 03-29-2022. https://integratedreporting.org/news/iirc-and-kirchhoff-consult-study-investors-rely-on-non-financial-information-corporate-reports-not-there-yet/.

Ingram, R. W., Brooks, L. D., and Copeland, R. M. 1983. “The Information Content of Municipal Bond Rating Changes: A Note.” *The Journal of Finance* 38 (3): 997-1003.

Irvine, P. J. 2004. “Analysts’ Forecasts and Brokerage-Firm Trading.” *The Accounting Review* 79 (1): 125-149.

Jackson, A. R. 2005. “Trade Generation, Reputation, and Sell-Side Analysts.” *The Journal of Finance* 60 (2): 673-717.

Jacobi, C., van Atteveldt, W., and Welbers, K. 2016. “Quantitative analysis of large amounts of journalistic texts using topic modelling.” *Digital Journalism* 4 (1): 89-106.

Jatnika, D., Bijaksana, M. A., and Suryani, A. A. 2019. “Word2Vec Model Analysis for Semantic Similarities in English Words.” *Procedia Computer Science* 157: 160-167.

Jensen, M. C., and Meckling, W. H. 1976. “Theory of the firm: Managerial behavior, agency costs and ownership structure.” *Journal of Financial Economics* 3 (4): 305-360.

Jianqiang, Z., and Xiaolin, G. 2017. “Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis.” *IEEE Access* 5: 2870-2879.

Jivani, A. G. 2011. “A comparative study of stemming algorithms.” *International Journal of Computer Applications in Technology* 2 (6): 1930-1938.

Johnsen, D. B. 2008. “The SEC's 2006 Soft Dollar Guidance: Law and Economics.” *Cardozo Law Review* 30 (4): 1545-1614.

Jonson, G., Woolf, P., and Bala, S. 2019. *CRH Feedback From Sell-Side Dinner with Mgmt.* Jeffries.

Jorden, J. F. 1975. “"Paying up" for Research: A Regulatory and Legislative Analysis.” *Duke Law Journal* 1975 (5): 1103-1129.

Jurafsky, D., and Martin, J. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. Upper Saddle River, United States: Prentice-Hall.

Kadhim, A. I. 2019. “Survey on supervised machine learning techniques for automatic text classification.” *The Artificial Intelligence Review* 52 (1): 273-292.

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., and Liu, H. 2020. “Natural language processing (NLP) in management research: A literature review.” *Journal of Management Analytics* 7 (2): 139-172.

Kearney, C., and Liu, S. 2014. “Textual sentiment in finance: A survey of methods and models.” *International Review of Financial Analysis* 33: 171-185.

Kempf, A., and Osthoff, P. 2007. “The Effect of Socially Responsible Investing on Portfolio Performance.” *European Financial Management* 13 (5): 908-922.

Kessel, L., and Graf-Vlachy, L. 2021. “Chief Digital Officers – Review and Research Agenda.” *Proceedings of the International Conference on Wirtschaftsinformatik*. Essen, Germany.

Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., and Visa, A. 2004. “Combining Data and Text Mining Techniques for Analysing Financial Reports.” *Intelligent Systems in Accounting, Finance and Management* 12 (1): 29-41.

Kothari, S. P., Shu, S., and Wysocki, P. D. 2009. “Do Managers Withhold Bad News?” *Journal of Accounting Research* 47 (1): 241-276.

Kotsantonis, S., and Serafeim, G. 2019. “Four Things No One Will Tell You About ESG Data.” *Journal of Applied Corporate Finance* 31 (2): 50-58.

KPMG. 2020. “The Time has Come: The KPMG Survey of Sustainability.” Accessed 03-29-2022. https://home.kpmg/xx/en/home/insights/2020/11/the-time-has-come-survey-of-sustainability-reporting.html.

Krakow, N. J., and Schäfer, T. 2020. “Mutual Funds and Risk Disclosure: Information Content of Fund Prospectuses.” Working Paper. Swiss Finance Institute Research Paper. Available at SSRN 3648089.

Kralina, L. 2018. “Big Data Signaling in Annual Reports: Does the Presence of a Chief Data Officer (CDO) Signal Big Data Initiatives in Annual Financial Reports?” *Journal of Business & Financial Affairs* 7 (2): 1-3.

Kuechler, B., and Vaishnavi, V. 2008. “On theory development in design science research: anatomy of a research project.” *European Journal of Information Systems* 17 (5): 489-504.

Kunisch, S., Menz, M., and Langan, R. 2022. “Chief digital officers: An exploratory analysis of their emergence, nature, and determinants.” *Long Range Planning* 55 (2): 1-17.

Kuperman, J. C., Athavale, M., and Eisner, A. 2003. “Financial analysts in the media: Evolving roles and recent trends.” *American Business Review* 21 (2): 74-80.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. 2015. “From Word Embeddings To Document Distances.” *Proceedings of the 32nd International Conference on Machine Learning*. Lille, France.

Lang, M., Pinto, J., and Sul, E. 2021. “MiFID II unbundling and sell side analyst research.” Working Paper. Available at SSRN: 3408198.

Lang, M., and Stice-Lawrence, L. 2015. “Textual analysis and international financial reporting: Large sample evidence.” *Journal of Accounting and Economics* 60 (2): 110-135.

Le, Q., and Mikolov, T. 2014. “Distributed Representations of Sentences and Documents.” *Proceedings of the 31st International Conference on Machine Learning*. Beijing, China.

Lechner, M. 2011. “The Estimation of Causal Effects by Difference-in-Difference Methods.” *Foundations and Trends in Econometrics* 4 (3): 165-224.

Lee, J., and Hong, Y. S. 2014. “Business Model Mining: Analyzing a Firm's Business Model with Text Mining of Annual Report.” *Industrial Engineering and Management Systems* 13 (4): 432-441.

Leonhardt, D., Hanelt, A., Huang, P., and Mithas, S. 2018. “Does One Size Fit All? Theorizing Governance Configurations for Digital Innovation.” *Proceedings of the International Conference on Information Systems*. San Francisco, United States.

Lev, B. 1988. “Toward a Theory of Equitable and Efficient Accounting Policy.” *The Accounting Review* 63 (1): 1-22.

Lewis, C., and Young, S. 2019. “Fad or future? Automated analysis of financial text and its implications for corporate reporting.” *Accounting and Business Research* 49 (5): 587-615.

Li, C. 2015. “The Hidden Face of the Media: How Financial Journalists Produce Information.” Working Paper.

Liu, R., Mai, F., Shan, Z., and Wu, Y. 2020. “Predicting Shareholder Litigation on Insider Trading from Financial Text: An Interpretable Deep Learning Approach.” *Information & Management* 57 (8): 103387.

Liu, Z., and Yezegel, A. 2020. “Was MiFID II effective in unbundling execution and research services?” Working Paper. Available at SSRN: 3525201.

Loughran, T., and McDonald, B. 2011. “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.” *The Journal of Finance* 66 (1): 35–65.

Loughran, T., and McDonald, B. 2016. “Textual Analysis in Accounting and Finance: A Survey.” *Journal of Accounting Research* 54 (4): 1187-1230.

Lu, Y. 2013. “Automatic topic identification of health-related messages in online health community using text classification.” *SpringerPlus* 2 (1): 309.

Luciano, M. M., Nahrgang, J. D., and Shropshire, C. 2020. “Strategic Leadership Systems: Viewing Top Management Teams and Boards of Directors from A Multiteam Systems Perspective.” *Academy of Management Review* 45 (3): 675-701.

Lys, T., and Sohn, S. 1990. “The association between revisions of financial analysts' earnings forecasts and security-price changes.” *Journal of Accounting and Economics* 13 (4): 341-363.

Mackie, W. 2015. *Power down.* Kepler Cheuvreux.

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., and Ye, L. 2005. “Author identification on the large scale.” *Proceedings of the 2005 Meeting of the Classification Society of North America*. St. Louis, United States.

Manning, C. D., Schütze, H., and Raghavan, P. 2008. *Introduction to information retrieval*. Cambridge, England: Cambridge University Press.

Matt, C., Hess, T., and Benlian, A. 2015. “Digital Transformation Strategies.” *Business & Information Systems Engineering* 57 (5): 339-343.

Mauritz, C., Nienhaus, M., and Oehler, C. 2021. “The role of individual audit partners for narrative disclosures.” *Review of Accounting Studies*.

McNichols, M., and Dravid, A. 1990. “Stock Dividends, Stock Splits, and Signaling.” *The Journal of Finance* 45 (3): 857-879.

Mehran, H., and Stulz, R. M. 2007. “The economics of conflicts of interest in financial institutions.” *Journal of Financial Economics* 85 (2): 267-296.

Melville, N. P. 2010. “Information Systems Innovation for Environmental Sustainability.” *MIS Quarterly* 34 (1): 1-21.

Meth, H., Mueller, B., and Maedche, A. 2015. “Designing a requirement mining system.” *Journal of the Association for Information Systems* 16 (9): 799-837.

Metzler, D. R., Bankamp, S., Muntermann, J., and Palmer, M. 2021. “The Role of CDOs in Signaling Digital Transformation Endeavors: An Analysis of Firms’ External Communication Tools.” *Proceedings of the International Conference on Information Systems*. Austin, United States.

Metzler, D. R., and Muntermann, J. 2020. “The Impact of Digital Transformation on Incumbent Firms: An Analysis of Changes, Challenges, and Responses at the Business Model Level.” *Proceedings of the International Conference on Information Systems*. Hyderabad, India.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013a. “Efficient Estimation of Word Representations in Vector Space.” Working Paper. Available at arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013b. “Distributed Representations of Words and Phrases and their Compositionality.” *Proceedings of the 26th International Conference on Neural Information Processing Systems*. Lake Tahoe, United States.

Miller, G. S., and Skinner, D. J. 2015. “The Evolving Disclosure Landscape: How Changes in Technology, the Media, and Capital Markets Are Affecting Disclosure.” *Journal of Accounting Research* 53 (2): 221-239.

Miller, M. H., and Rock, K. 1985. "Dividend Policy under Asymmetric Information." *The Journal of Finance* 40 (4): 1031-1051.

Moker, A., Brosi, P., and Welpe, I. M. 2020. "It Depends on the Size: How Firm Strategic Emphasis on Digital Transformation Predicts Market Capitalization." *Proceedings of the Hawaii International Conference on System Sciences*. Maui, United States.

Mola, S., Rau, P. R., and Khorana, A. 2013. "Is There Life after the Complete Loss of Analyst Coverage?" *The Accounting Review* 88 (2): 667-705.

Morris, S. B. 2008. "Estimating Effect Sizes From Pretest-Posttest-Control Group Designs." *Organizational Research Methods* 11 (2): 364-386.

Myers, S. C., and Majluf, N. S. 1984. "Corporate financing and investment decisions when firms have information that investors do not have." *Journal of Financial Economics* 13 (2): 187-221.

Naseem, U., Razzak, I., and Eklund, P. W. 2020. "A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter." *Multimedia Tools and Applications* 80 (28): 35239-35266.

Nilsson, H., Cunningham, G. M., and Hassel, L. G. 2008. "A study of the provision of environmental information in financial analysts' research reports." *Sustainable Development* 16 (3): 180-194.

Niraula, N., Banjade, R., Ştefănescu, D., and Rus, V. 2013. "Experiments with Semantic Similarity Measures Based on LDA and LSA." In *Statistical Language and Speech Processing. SLSP 2013.*, edited by Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan Mitkov and Bianca Truthe, 188-199. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg.

Nofsinger, J., and Varma, A. 2014. "Socially responsible funds and market crises." *Journal of Banking & Finance* 48: 180-193.

Oshikawa, R., Qian, J., and Wang, W. Y. 2020. "A Survey on Natural Language Processing for Fake News Detection." *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France.

OuYang, Z., Xu, J., Wei, J., and Liu, Y. 2017. "Information Asymmetry and Investor Reaction to Corporate Crisis: Media Reputation as a Stock Market Signal." *Journal of Media Economics* 30 (2): 82-95.

Palmer, M., Eickhoff, M., and Muntermann, J. 2018. "Detecting Herding Behavior Using Topic Mining: The Case of Financial Analysts." *Proceedings of the 26th European Conference on Information Systems*. Portsmouth, United Kingdom.

Pástor, Ľ., Stambaugh, R. F., and Taylor, L. A. 2021. "Sustainable investing in equilibrium." *Journal of Financial Economics* 142 (2): 550-571.

Pavlou, P. A., Liang, H., and Xue, Y. 2007. "Understanding and Mitigating Uncertainty in Online Exchange Relationships: A Principal-Agent Perspective." *MIS Quarterly* 31 (1): 105-136.

Payne, J. L., and Thomas, W. B. 2003. "The Implications of Using Stock-Split Adjusted I/B/E/S Data in Empirical Research." *The Accounting Review* 78 (4): 1049-1067.

Péladeau, P., Herzog, M., and Acker, O. 2017. "The New Class of Digital Leaders." *Strategy + Business* 88.

Pencle, N., and Mălăescu, I. 2016. "What's in the Words? Development and Validation of a Multidimensional Dictionary for CSR and Application Using Prospectuses." *Journal of Emerging Technologies in Accounting* 13 (2): 109-127.

Pennington, J., Socher, R., and Manning, C. D. 2014. "GloVe: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar.

Petersen, M. A. 2009. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches." *The Review of Financial Studies* 22 (1): 435-480.

Pope, P. F., Tamayo, A. M., and Wang, Y. H. 2019. "The impact of separating sell-side research payments from dealing commissions: Evidence from Sweden." Working Paper.

Porter, M. F. 1980. "An algorithm for suffix stripping." *Program* 14 (3): 130-137.

Pradhan, N., Gyanchandani, M., and Wadhvani, R. 2015. "A Review on Text Similarity Technique used in IR and its Application." *International Journal of Computer Applications* 120 (9): 29-34.

Preece, R. 2017. "MiFID II: A New Paradigm for Investment Research." CFA Institute. Accessed 03-29-2022. https://www.cfainstitute.org/en/research/survey-reports/mifid-ii-a-new-paradigm-for-investment-research.

Preece, R. 2019. "MiFID II: One Year On." CFA Institute. Accessed 03-29-2022. https://www.cfainstitute.org/en/research/survey-reports/mifid-II-one-year-on.

PwC. 2020. "The growth opportunity of the century." Sustainable Finance Series. Accessed 03-29-2022. https://www.pwc.lu/en/sustainable-finance/docs/pwc-esg-report-the-growth-opportunity-of-the-century.pdf.

Ramakrishnan, R. T. S., and Thakor, A. V. 1984. "Information Reliability and a Theory of Financial Intermediation." *The Review of Economic Studies* 51 (3): 415-432.

RavenPack. 2021. "RavenPack Products." Accessed 10-13-2021. https://www.ravenpack.com/products?news-analytics-tab.

Rawte, V., Gupta, A., and Zaki, M. J. 2021. "A Comparative Analysis of Temporal Long Text Similarity: Application to Financial Documents." In *Mining Data for Financial Applications. MIDAS 2020. Lecture Notes in Computer Science*, edited by Valerio Bitetta, Ilaria Bordino, Andrea Ferretti, Francesco Gullo, Giovanni Ponti and Lorenzo Severini, In Bitetta V., Bordino I., Ferretti A., Gullo F., Ponti G., Severini L. (eds) Mining Data for Financial Applications. MIDAS 2020. Lecture Notes in Computer Science, 77-91. Cham, Switzerland: Springer International Publishing.

Reck, F., and Fliaster, A. 2018. "How to Drive Digital Transformation? – A Configurational Analysis on the Impact of CDOs." *Academy of Management Annual Meeting Proceedings*. Chicago, United States.

Reck, F., and Fliaster, A. 2019. "Four profiles of successful digital executives." *MIT Sloan Management Review* 60 (3): 1-7.

Rees, L., Sharp, N., and Twedt, B. 2015. "Who's heard on the Street? Determinants and consequences of financial analyst coverage in the business press." *Review of Accounting Studies* 20 (1): 173-209.

Refinitiv. 2022. "The Refinitiv Business Classification Methodology." Accessed 05-19-2022. https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/trbc-business-classifcation-methodology.pdf.

Rhee, R. J. 2015. "Why Credit Rating Agencies Exist." *Economic Notes* 44 (2): 161-176.

Richardson, V. J. 2000. "Information Asymmetry and Earnings Management: Some Evidence." *Review of Quantitative Finance and Accounting* 15 (4): 325-347.

Rose, F. 1999. *The Economics, Concept, and Design of Information Intermediaries*. Heidelberg, Germany: Physica-Verlag.

Rosenbaum, P. R., and Rubin, D. B. 1983. “The central role of the propensity score in observational studies for causal effects.” *Biometrika* 70 (1): 41-55.

Şaşmaz, E., and Tek, F. B. 2021. “Tweet Sentiment Analysis for Cryptocurrencies.” *6th International Conference on Computer Science and Engineering (UBMK)*. Ankara, Turkey.

Schmalensee, R. 1984. “Gaussian Demand and Commodity Bundling.” *The Journal of Business* 57 (1): 211-230.

Schoonmaker, K. 2018. *Caterpillar Grows 2Q Top Line an Impressive 24%, Attains Record 2Q EPS.* Morningstar Equity Research.

Schueth, S. 2003. “Socially Responsible Investing in the United States.” *Journal of Business Ethics* 43 (3): 189-194.

Schweiger, D. M., and Denisi, A. S. 1991. “Communication with Employees Following a Merger: A Longitudinal Field Experiment.” *Academy of Management Journal* 34 (1): 110-135.

Sebastian, I. M., Ross, J. W., Beath, C. M., Mocker, M., Moloney, K., and Fonstad, N. O. 2017. “How Big Old Companies Navigate Digital Transformation.” *MIS Quarterly Executive* 16 (3): 197-213.

SEC. 2017. “Investment Advisers Act of 1940 - Section 202(a)(11) Securities Industry and Financial Markets Association.” Accessed 03-29-2022. http://www.sec.gov/divisions/investment/noaction/2017/sifma-102617-202a.htm.

SEC. 2019. “Investment Advisers Act of 1940 - Section 202(a)(11) Securities Industry and Financial Markets Association.” Accessed 03-29-2022. https://www.sec.gov/investment/sifma-110419.

SEC. 2021. “Form 10-K.” U.S. Securities and Exchange Commission (SEC). Accessed 04-28-2021. https://www.investor.gov/introduction-investing/investing-basics/glossary/form-10-k/.

Shahmirzadi, O., Lugowski, A., and Younge, K. 2019. “Text Similarity in Vector Space Models: A Comparative Study.” *2019 18th IEEE International Conference On Machine Learning And Applications*. Boca Raton, United States.

Shi, Y., Liu, W.-M., and Ho, K.-Y. 2016. “Public news arrival and the idiosyncratic volatility puzzle.” *Journal of Empirical Finance* 37: 159-172.

Siering, M., Muntermann, J., and Grčar, M. 2021. “Design Principles for Robust Fraud Detection: The Case of Stock Market Manipulations.” *Journal of the Association for Information Systems* 22 (1): 156-178.

Silva, C., and Ribeiro, B. 2003. “The importance of stop word removal on recall values in text categorization.” *Proceedings of the International Joint Conference on Neural Networks*. Portland, United States.

Singh, A., and Hess, T. 2017. “How Chief Digital Officers Promote the Digital Transformation of their Companies.” *MIS Quarterly Executive* 16 (1): 1-17.

Singh, A., Klarner, P., and Hess, T. 2020. “How do chief digital officers pursue digital transformation activities? The role of organization design parameters.” *Long Range Planning* 53 (3): 101890.

Singhal, A. 2001. “Modern Information Retrieval: A Brief Overview.” *IEEE Data Engineering Bulletin* 24 (4): 35-43.

Sparck Jones, K. 1972. “A statistical interpretation of term specificity and its application in retrieval.” *Journal of Documentation* 28 (1): 11-21.

Spence, M. 1973. “Job Market Signaling.” *The Quarterly Journal of Economics* 87 (3): 355-374.

Spence, M. 2002. “Signaling in Retrospect and the Informational Structure of Markets.” *The American Economic Review* 92 (3): 434-459.

Spindler, G., and Jaenig, R. 2010. “Informing Shareholders and Investors: A Behavioral and Economics Approach from a German Company Law Perspective.” *Journal of Interdisciplinary Economics* 22 (1-2): 89-108.

Stiglitz, J. E. 1975. “The Theory of "Screening," Education, and the Distribution of Income.” *The American Economic Review* 65 (3): 283-300.

Stolper, A. 2009. “Regulation of credit rating agencies.” *Journal of Banking & Finance* 33 (7): 1266-1273.

Stone, P. J., Dunphy, D. C., and Smith, M. S. 1966. *The general inquirer: A computer approach to content analysis*. Cambridge, United States: M.I.T. Press.

Su, W., Peng, M. W., Tan, W., and Cheung, Y.-L. 2016. “The Signaling Effect of Corporate Social Responsibility in Emerging Economies.” *Journal of Business Ethics* 134 (3): 479-491.

Sun, C., Qiu, X., Xu, Y., and Huang, X. 2019. “How to Fine-Tune BERT for Text Classification?” *China National Conference on Chinese Computational Linguistics*. Kunming, China.

Tambini, D. 2010. “What are financial journalists for?” *Journalism Studies* 11 (2): 158-174.

Tan, H., Wang, S., and Welker, M. 2011. “Analyst Following and Forecast Accuracy After Mandated IFRS Adoptions.” *Journal of Accounting Research* 49 (5): 1307-1357.

Tan, M., Dos Santos, C., Xiang, B., and Zhou, B. 2016. “Improved Representation Learning for Question Answer Matching.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany.

Trieu, L. Q., Tran, H. Q., and Tran, M.-T. 2017. “News Classification from Social Media Using Twitter-based Doc2Vec Model and Automatic Query Expansion.” *Proceedings of the Eighth International Symposium on Information and Communication Technology*. Nha Trang City, Vietnam.

Trueman, B. 1994. “Analyst Forecasts and Herding Behavior.” *The Review of Financial Studies* 7 (1): 97-124.

Tumbas, S., Berente, N., and vom Brocke, J. 2017. “Three Types of Chief Digital Officers and the Reasons Organizations Adopt the Role.” *MIS Quarterly Executive* 16 (2): 121-134.

Tumbas, S., Berente, N., and vom Brocke, J. 2018. “Digital Innovation and Institutional Entrepreneurship: Chief Digital Officer Perspectives of their Emerging Role.” *Journal of Information Technology* 33: 188-202.

Uang, J.-Y., Citron, D. B., Sudarsanam, S., and Taffler, R. J. 2006. “Management Going-concern Disclosures: Impact of Corporate Governance and Auditor Reputation.” *European Financial Management* 12 (5): 789-816.

UN General Assembly. 2015. “Transforming our World: the 2030 Agenda for Sustainable Development.” Accessed 03-29-2022. https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E.

Unruh, G., Kiron, D., Kruschwitz, N., Reeves, M., Rubel, H., and Zum Felde, A. M. 2016. “Investing for a sustainable future: Investors care more about sustainability than many executives believe.” *MIT Sloan Management Review* 57 (4): 3-29.

van Duuren, E., Plantinga, A., and Scholtens, B. 2016. “ESG Integration and the Investment Management Process: Fundamental Investing Reinvented.” *Journal of Business Ethics* 138 (3): 525-533.

Vasilescu, L., Hwang, W., Wakeham, S., Huang, L., Chang, T., and Jian, C. 2018. *The Moonshot: How Nike Changes its Footprint.* Macquarie Research.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. "Attention Is All You Need." *31st Conference on Neural Information Processing Systems*. Long Beach, United States.

Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Qi Dong, J., Fabian, N., and Haenlein, M. 2021. "Digital transformation: A multidisciplinary reflection and research agenda." *Journal of Business Research* 122: 889-901.

Vermaelen, T. 1981. "Common stock repurchases and market signalling: An empirical study." *Journal of Financial Economics* 9 (2): 139-183.

Vial, G. 2019. "Understanding digital transformation: A review and a research agenda." *The Journal of Strategic Information Systems* 28 (2): 118-144.

Viswanathan, S., Kuruzovich, J., Gosain, S., and Agarwal, R. 2007. "Online Infomediaries and Price Discrimination: Evidence from the Automotive Retailing Sector." *Journal of Marketing* 71 (3): 89-107.

Walsh, P. R., Webb, C. L., Thrum, A., and Rajan, N. 2018. *Solving for the CO2 problem.* Morgan Stanley.

Wang, J. L., Chung, C.-F., and Tzeng, L. Y. 2008. "An Empirical Analysis of the Effects of Increasing Deductibles on Moral Hazard." *Journal of Risk and Insurance* 75 (3): 551-566.

Webster, J. J., and Kit, C. 1992. "Tokenization as the initial phase in NLP." *Proceedings of the 14th conference on Computational linguistics*. Nantes, France.

Westerman, G., Bonnet, D., and McAfee, A. 2014. *Leading digital: Turning technology into business transformation*. Boston, United States: Harvard Business Review Press.

Womack, R. 2002. "Information intermediaries and optimal information distribution." *Library & Information Science Research* 24 (2): 129-155.

Wooldridge, J. M. 2015. *Introductory Econometrics: A Modern Approach*. 6 ed. Boston, United States: Cengage Learning.

Yan, J., Wang, K., Liu, Y., Xu, K., Kang, L., Chen, X., and Zhu, H. 2018. "Mining social lending motivations for loan project recommendations." *Expert Systems with Applications* 111: 100-106.

Yang, R., Yu, Y., Liu, M., and Wu, K. 2018. "Corporate Risk Disclosure and Audit Fee: A Text Mining Approach." *European Accounting Review* 27 (3): 583-594.

Yang, Y., UY, M. C. S., and Huang, A. 2020. "FinBERT: A Pretrained Language Model for Financial Communications." Working Paper. Available at arXiv: 2006.08097.

Zahedi, F. M., and Song, J. 2008. "Dynamics of Trust Revision: Using Health Infomediaries." *Journal of Management Information Systems* 24 (4): 225-248.

Zhan, X., and Mu, Y. 2016. "Examining the shareholder value effects of announcements of CDO positions." *13th International Conference on Service Systems and Service Management (ICSSSM)*. Kunming, China.

Zhang, Y., Callan, J., and Minka, T. 2002. "Novelty and Redundancy Detection in Adaptive Filtering." *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. Tampere, Finland.

Zhang, Y., and Wiersema, M. F. 2009. "Stock market reaction to CEO certification: the signaling role of CEO background." *Strategic Management Journal* 30 (7): 693-710.

Zhao, R., and Mao, K. 2018. “Fuzzy Bag-of-Words Model for Document Representation.” *IEEE Transactions on Fuzzy Systems* 26 (2): 794-804.

Zimmerman, M. A. 2008. “The Influence of Top Management Team Heterogeneity on the Capital Raised through an Initial Public Offering.” *Entrepreneurship Theory and Practice* 32 (3): 391-414.

Appendix

Author contribution to individual studies of this thesis

Study	Conference	Status	Authors	Contribution
Bankamp and Muntermann (2022) Paper I.1	PACIS 2022	Published	**Bankamp** Muntermann	**95%** 5%
Bankamp and Muntermann (2021) Paper II.1	PACIS 2021	Published	**Bankamp** Muntermann	**90%** 10%
Bankamp et al. (2022) Paper II.2	INQUIRE Autumn Residential Seminar	Presented	**Bankamp** Palmer Muntermann	**80%** 10% 10%
Bankamp (2022) Paper II.3	12th Annual Pre-ICIS Workshop on Accounting Information Systems	Presented	**Bankamp**	**100%**
	44th Annual Congress of the European Accounting Association	Presented		
Metzler et al. (2021) Paper II.4	ICIS 2021	Published	Metzler **Bankamp** Muntermann Palmer	45% **45%** 5% 5%

19.06.2022, ______________________________

Datum, Unterschrift

Promotionsstudiengang „Wirtschaftswissenschaften“

Versicherung bei Zulassung zur Promotionsprüfung

Ich versichere,

1. dass ich die eingereichte Dissertation **Information Asymmetries and the Role of Information Intermediaries on Capital Markets s**elbstständig angefertigt habe und nicht die Hilfe Dritter in einer dem Prüfungsrecht und wissenschaftlicher Redlichkeit widersprechenden Weise in Anspruch genommen habe,
2. dass ich das Prüfungsrecht einschließlich der wissenschaftlichen Redlichkeit – hierzu gehört die strikte Beachtung des Zitiergebots, so dass die Übernahme fremden Gedankenguts in der Dissertation deutlich gekennzeichnet ist – beachtet habe,
3. dass beim vorliegenden Promotionsverfahren kein Vermittler gegen Entgelt eingeschaltet worden ist sowie im Zusammenhang mit dem Promotionsverfahren und seiner Vorbereitung
   - kein Entgelt gezahlt oder entgeltgleiche Leistungen erbracht worden sind
   - keine Dienste unentgeltlich in Anspruch genommen wurden, die dem Sinn und Zweck eines Prüfungsverfahrens widersprechen
4. dass ich eine entsprechende Promotion nicht anderweitig beantragt und hierbei die eingereichte Dissertation oder Teile daraus vorgelegt habe.

Mir ist bekannt, dass Unwahrheiten hinsichtlich der vorstehenden Versicherung die Zulassung zur Promotionsprüfung ausschließen und im Falle eines späteren Bekanntwerdens die Promotionsprüfung für ungültig erklärt werden oder der Doktorgrad aberkannt werden kann.

19.06.2022, ______________________________

Datum, Unterschrift